Universidad
Carlos III de Madrid
www.uc3m.es

DEPARTAMENTO DE TEORÍA DE LA SEÑAL Y COMUNICACIONES

TRABAJO FIN DE GRADO

Título: Localización en interiores mediante captura y clasificación de medidas reales de potencia de señal

Autor: Ignacio Gallego Yuste

Titulación: Grado en Ingeniería de Sistemas de Comunicaciones

Profesor: Miguel Lázaro Gredilla

Fecha: Marzo 2013

Título: Localización en interiores mediante captura y clasificación de medidas reales de potencia de señal
Autor: Ignacio Gallego Yuste
Tutor: Miguel Lázaro Gredilla

EL TRIBUNAL

Presidente:

Vocal:

Secretario:

Una vez terminada la lectura y defensa del trabajo fin de grado (TFG) en el mes de Marzo del año 2013, en la Escuela Politécnica Superior de la Universidad Carlos III de Madrid (Leganés), la calificación otorgada es :

SECRETARIO VOCAL PRESIDENTE

Agradecimientos

Después de unos meses de duro trabajo, por fin llega la recompensa esperada, que no es otra de lograr la satisfacción de terminar un ciclo de estudios. Pero a este punto no podría haber llegado sin la ayuda de muchas personas a las que quiero agradecer su apoyo.

En primer lugar, quiero agradecer a Miguel Lázaro Gredilla el haberme concedido la oportunidad de desarrollar este proyecto que ha resultado bastante interesante. Siempre se ha adaptado a mis circunstancias y hemos podido estar en permanente contacto. Además siempre ha estado al día con el proyecto, lo que hizo mucho más fácil la resolución de dudas.

A mis padres por poder permitirme estudiar la carrera, que con los tiempos que corren, se puede decir que es un lujo. Parece mentira que haya que decir esto.

A mis hermanos, por aconsejarme en las dudas que me han surgido en cada momento y siempre desear lo mejor para mí.

A Laura, por aguantarme desde hace año y medio, siendo un importantísimo apoyo desde entonces que, día a día, ha conseguido que sea más feliz.

A todos los compañeros y amigos de la universidad, pero sobre todos los amigos, que son los que después de que todo esto acabe, los únicos a los que seguiré viendo.

No hace falta nombrarlos porque ellos saben quiénes son.

A todos, mil gracias por hacer de esta época de mi vida la mejor que he vivido hasta el momento.

Resumen

Actualmente vivimos en una sociedad cuya demanda tecnológica está alcanzando límites impensables no hace tanto tiempo. Equipar un coche para hacer un viaje ya no consiste únicamente en preparar las maletas y hacer los bocadillos. Aparatos como el GPS, móvil, Dvd portátil, mp3, tableta… se han convertido en tan importantes como llevar una muda limpia.

Múltiples compañías han sido capaces de crear necesidades que no existían en la sociedad con la aparición de sus productos. Libros electrónicos, tablets, portafotos electrónicos… innumerables productos de dudosa efectividad se han colado en la espiral de consumo tecnológico en la que estamos implantados.

Pero dichos productos no han tenido éxito sin más, sino que detrás hay un trabajo de varios ingenieros (y todo su equipo) que han sido capaces de no solo crear el producto sino de hacerlo atractivo para la sociedad, independientemente del uso que esta pueda llegar a darle.

Y aquí nos encontramos nosotros, con el objetivo de crear una herramienta de posicionamiento indoor, las cuales están comenzando a gozar de una gran popularidad y están extendiendo su presencia en el mercado, pero que aún tienen que perfeccionar muchos aspectos de su funcionamiento para ofrecer un servicio impecable.

El objetivo que queremos alcanzar es que se pueda determinar de manera fiable la estancia donde se encuentra una persona, un objeto, un aparato…..en una superficie cerrada como pueda ser un centro comercial, un hangar o un teatro.

Para ello hemos recopilado multitud de medidas reales de potencia de señal WIFI en múltiples estancias, tanto en la universidad como en una vivienda unifamiliar. Nos hemos servido, principalmente, de un Smartphone (HTC DESIRE) y dos aplicaciones (WIFI Analyzer y Salamander) para la creación de una base de datos, que nos servirá para el entrenamiento de los algoritmos de clasificación multiclase que hemos empleado para la clasificación de las muestras.

Dicha base de datos ha sido conformada mediante la extracción de los datos de las aplicaciones empleadas en diversos ficheros y posteriormente moldeados hasta conseguir una matriz de datos que se pudiera tratar cómodamente. Una vez llegados a esta situación, con los datos recopilados y presentados a la vista de manera agradable, entra en acción la herramienta matemática por excelencia: Matlab.

Una vez tenemos los datos cargados en Matlab, observaremos las tasas de acierto que obtendremos de los diferentes algoritmos de clasificación multiclase empleados, dependiendo de dichas tasas y del tiempo empleado tanto en el entrenamiento de los algoritmos con en la clasificación de las muestras, analizaremos los algoritmos más convenientes para cada ocasión.

Abstract

Nowadays we live in a society whose technological requirements are reaching limits which were unbelievable not much time ago. Packing for a road trip is no longer as easy as putting some food and a couple of bottles of water. Devices like GPS, mobile phones, tablet, mp3… have become as necessary bringing clean underwear.

A lot of companies have been able of create necessities that never existed before in our society with the creation of their products. Electronic books, tablets, electronic photo frames, etc., are examples of products that despite their dubious effectiveness have managed to get ourselves into a spiral of consumption that does not seem to be leaving us any soon.

In this work we set out to create a new indoor positioning tool. Positioning tools are beginning to enjoy great popularity and are expanding their presence in the market, but many aspects of its operation are yet to be improved in order to offer an impeccable service.

The goal that we want to achieve is to be able to reliably determine in which area a given person, object, or device are located when they lie inside a closed space, such as malls, hangars or theaters.

In order to do this, we have recorded a lot of measurements of WIFI signal strength in multiples areas, like the university campus or a family house. We have used a Smartphone (HTC Desire) and two applications (WIFI Analyzer and Salamander) for the creation of a database that we will use for training different multiclass classification algorithms.

The database has been populated from the data that the applications have collected in different areas. Once we have reached this point of the work and the collected data is pleasantly presented to the eye, it is time for some mathematical modeling using Matlab.

Once the data are loaded in Matlab, we will apply different multiclass classification algorithms in order to determine which of them is the best suited for our objective, according to performance measurements of accuracy and speed.

ÍNDICE

1. Motivación

El GPS ha subsanado la mayor parte de los problemas de localización existentes en entornos exteriores instaurándose en la vida cotidiana de la mayoría de las personas. El desarrollo que ha experimentado este sistema en estos últimos años ha provocado que su coste sea soportable por la mayoría de los bolsillos, añadiendo a su favor que después de su adquisición, su uso no supone más costes económicos adicionales.

Su éxito se refleja en las innumerables aplicaciones que este sistema posee tanto en el ámbito civil y profesional como en el ámbito militar. Pero hay un obstáculo que el GPS aún no ha podido sobrepasar, su funcionamiento en espacios interiores.

Es aquí donde se centra nuestro proyecto, en desarrollar una herramienta capaz de dar un servicio de posicionamiento interior (indoor), precisamente donde el GPS no puede proporcionarlo en la actualidad.

Para llevar a cabo este proyecto nos hemos servido de la tecnología que nos proporcionan las redes inalámbricas, en concreto de la señal WIFI, debido a la facilidad para medir la potencia de dichas señales.

Además la expansión de este tipo de herramientas resulta bastante sencilla ya que los teléfonos móviles de última generación (Smartphone) cada vez están más expandidos y pueden usarse como plataformas para la instalación de la aplicación al estar preparados para la captación de señales WIFI.

Visualicemos por un momento la utilidad que esta aplicación podría tener:

Nos encontramos en algún centro comercial de superficie extensa o en unos grandes almacenes de varias plantas. Allí observamos carteles por todos los lados que tarde o temprano nos acabarán llevando al restaurante, tiendas o cines que buscamos.

Pero imagínense que, basándonos en la potencia de la señal WIFI emitida por los distintos puntos de acceso, nuestra herramienta sería capaz de indicarnos la estancia donde nos encontramos indicándonos las tiendas y restaurantes que nos rodean, pudiendo visualizar cómodamente, en la pantalla del dispositivo, información útil para el consumidor.

Otro ejemplo en el que su uso puede parecer aún más eficaz es el caso de las personas invidentes. La persona invidente, con esta herramienta instalada en su dispositivo móvil, podría saber en cada instante en que estancia de una estación de metro se encuentra. Para ello, nuestra herramienta se basaría en la potencia de señal que el dispositivo móvil recibe de los distintos puntos de acceso de la estación. Simplemente añadiendo la funcionalidad de voz a la aplicación, esta reproduciría por el altavoz del dispositivo el nombre de la estancia.

Hemos ilustrado con dos aplicaciones bastante representativas el servicio que nuestra herramienta podría proporcionar, pero son múltiples los ámbitos en los que el posicionamiento indoor puede ser útil.

A pesar de que en los últimos años el desarrollo de este tipo de herramientas está en auge, aún queda mucho por hacer y por ello vamos a aportar nuestro granito de arena.

El objetivo final que queremos alcanzar es el de desarrollar una herramienta que, basándose en los datos reales recopilados de potencia de señal WIFI con los que posteriormente hemos entrenado a una serie de algoritmos, sea capaz de averiguar la estancia donde se encuentra cualquier persona u objeto analizando la potencia de señal WIFI que su dispositivo móvil está recibiendo de los distintos puntos de acceso a la red Wireless.

2. Estado del arte

Antes de entrar de lleno en la explicación del proyecto que tenemos entre manos, vamos a poner en contexto y situar al TFG respecto a la evolución de las comunicaciones inalámbricas, sistemas de posicionamiento global y sistemas de posicionamiento en interiores.

2.1 Historia de las comunicaciones inalámbricas

2.1.1 Origen

Si queremos hablar de la historia de las redes inalámbricas debemos remontarnos a 1880 coincidiendo con la invención del primer aparato de comunicación sin cables, el fotófono.

Sus inventores fueron Graham Bell y Summer Tainter. No tuvo mucho éxito debido a que el fotófono permitía la transmisión del sonido por medio de emisión de luz y en aquella época aún no se distribuía la electricidad.

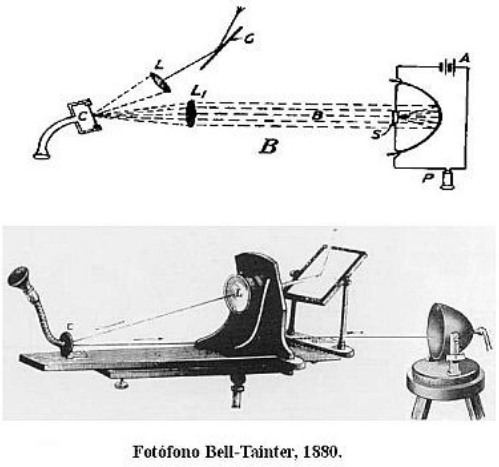


Figura1. Fotófono desarrollado por Bell- Tainter. [1]

El siguiente avance en la comunicación sin cables data de 1888 y fue llevado a cabo por el físico alemán Rudolf Hertz. Consiguió realizar una transmisión sin cables con ondas electromagnéticas mediante un oscilador y un resonador que fueron usados como emisor y receptor respectivamente.

Fue en 1899 cuando Guillermo Marconi estableció comunicaciones inalámbricas a través del Canal de la Mancha y en 1907 se transmitían los primeros mensajes completos a través del Atlántico.

La Segunda Guerra Mundial dio pie a numerosos avances en este campo.

Ya en 1971, en la Universidad de Hawaii, se creó el primer sistema de conmutación de paquetes mediante una red de comunicación por radio conocida como ALOHA. Estaba formada por 7 ordenadores situados en las distintas islas del archipiélago y eran capaces de comunicarse con un servidor central. Este primer sistema es conocido como la primera red de área local inalámbrica (WLAN).

2.1.2 Funcionamiento

Con el objetivo de transportar información de un punto a otro a través de ondas electromagnéticas, se hace uso de ondas portadoras y ondas moduladoras.

Las ondas portadoras son de una frecuencia mucho más alta que las moduladoras, que son las que contienen la información a transmitir. La modulación se produce cuando se realiza el acoplo de la onda moduladora con la portadora. Dependiendo de las técnicas de modulación existen una o varias portadoras. En las tres técnicas de modulación básica solo se utiliza una, y estas son:

☑ Modulación de la amplitud (AM)

☑ Modulación de la fase (PM)

☑ Modulación frecuencia (FM)

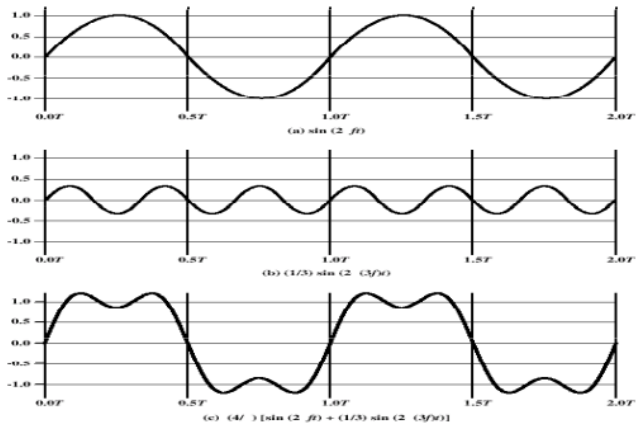


Figura 2. Ondas portadoras y ondas moduladas [1]

2.1.3 Estándares

La compatibilidad entre equipos era uno de los grandes problemas que podrían evitar la expansión de las redes inalámbricas. Así pues, la IEEE creó un grupo de trabajo específico para esta tarea de compatibilidades llamado 802.11.

El problema no se llegó a solventar del todo hasta 1996 con la creación del organismo Wireless Fidelity Alliance (WIFI Alliance), ya que el estándar 802.11 constaba de varios puntos con cierta ambigüedad que podían ser interpretados de distinta manera por los fabricantes provocando incompatibilidades entre equipos.

La WIFI Alliance permitió homogeneizar productos y hacer posible su implantación en el mercado de consumo. A partir de aquí fue cuando estas redes comenzaron a ser usadas por el público en general.

Fue tal la labor que desempeño la WIFI Alliance que hoy en día, las redes inalámbricas se conocen como redes WIFI en referencia a este organismo.

2.1.3.1 802.11 b

Dentro del seno de la IEEE eran conscientes de los problemas de ambigüedad generados por el estándar 802.11. Por ello, en 1999 surgió la norma 802.11 b como una evolución de la 802.11.

Con la adopción de esta nueva norma la popularidad de la redes WIFI creció exponencialmente pues la velocidad que ofrecía, a pesar de estar lejos de la proporcionada por la red cableada, era la suficiente para realizar las tareas más comunes. Trabaja en la banda de 2,4 GHz llegando a alcanzar velocidades de hasta 11 Mb/s.

2.1.3.2 802.11 a

La aprobación de la norma 802.11 a también data de 1999 pero diversos problemas provocaron una adopción muy lenta, particularmente en Europa.

Ofrece una velocidad máxima de 54 Mb/s y establece la banda de 5 GHz para el funcionamiento de los equipos. El mayor inconveniente se produjo cuando en el momento de su aprobación, esa banda del espectro electromagnético Europeo estaba asignada a usos privados. No fue hasta el año 2002 cuando la banda de los 5 GHz se liberalizó permitiendo el uso de estos equipos en Europa.

Pero ya en esas fechas la base implantada de sistemas con la norma 802.11 b hacía muy costosa la migración hacia 802.11 a debido a que no son compatibles entre sí. Su retraso en la salida al mercado, menor número de sistemas compatibles y el coste de reciclado de equipos fueron obstáculos muy difíciles de solventar para adopción de esta nueva norma.

2.1.3.3 802.11 g

Aprobada en el año 2003, la norma 802.11 g cuenta con dos ventajas principales:

☑ **Velocidad**: puede alcanzar una velocidad máxima de 54 Mb/s, lo que supone prácticamente quintuplicar la velocidad de la norma 802.11 b.
☑ **Compatibilidad:** es perfectamente compatible con la base de equipos WIFI que se rigen bajo la norma 802.11.b.

Dicha compatibilidad no es del todo perfecta y vamos a explicar por qué.

Generalmente las redes WIFI están compuestas por varias celdas donde podemos encontrar usuarios tanto de la norma b como de la g. Esto provoca que todos los puntos de acceso de la red hayan de funcionar en modo compatible b/g debido a que han de dar servicio a los clientes con movilidad que pasan de una zona de cobertura a otra. Esto provoca un empeoramiento de las condiciones de la red pues deben de utilizarse tiempos más lentos y compatibles con 802.11 b disminuyendo la velocidad ofrecida por la celda al trabajar en modo dual.

De manera que resulta muy recomendable fijar el modo puro 802.11 g y evitar el modo de compatibilidad b/g en las redes que no se prevea dar servicio a clientes 802.11 b.

2.1.3.4 802.11 n

Fue ya en el 2007 cuando la norma 802.11 n se publicó con el objetivo de proporcionar una mayor velocidad pasando de los 54 Mb/s a unos teóricos 600 Mb/s. Hablamos de teóricos 600 Mb/s porque lo normal son unos 300Mb/s a excepción de sistemas que se sitúan en torno a los 450 Mb/s, sin ser lo habitual.

Esta norma nos ofrece la posibilidad de funcionar en ambas bandas, tanto en 2,4 GHz como en la de 5 GHz. La compatibilidad con las normas anteriores es una clara ventaja para la integración de nuevos sistemas en redes ya existentes.

Actualmente la norma 802.11 n funcionando en la banda de los 2,4 GHz es donde encontraremos mayores ofertas de sistemas en el mercado. Sistemas para esta norma que funcionen en modo dual o en los 5 GHz apenas encontraremos. Esta tendencia se instauró en el mercado debido al menor coste y la compatibilidad con los sistemas anteriores que funcionan en la banda de 2,4 GHz.

Al igual que observamos en la norma 802.11 g la compatibilidad con el resto de normas no es perfecta y también experimentaremos una reducción significativa de la velocidad de transmisión en una red si la obligamos a trabajar en modo dual al contener clientes de distintas normas.

2.2 Historia de los sistemas de posicionamiento global

El GPS (sistema de posicionamiento global) es un sistema de navegación creado por el departamento de defensa de los EEUU, compuesto por un conjunto de satélites puestos en órbita y por una serie de estaciones receptoras en tierra firme.

Para determinar automáticamente la posición que ocupamos en la tierra (latitud y longitud) debemos de usar el GPS. Funciona continuamente en todas las partes del mundo y sólo necesitamos disponer de un receptor GPS sin más costes adicionales.

Sus orígenes militares no han impedido que actualmente formen parte nuestra vida cotidiana.

2.2.1 Orígenes del GPS

El desarrollo de la tecnología GPS surgió debido a la necesidad de las fuerzas armadas de poseer un sistema de navegación de alta precisión y que estuviera disponible para diversas aplicaciones.

Su desarrollo radica en los estudios de dispositivos extremadamente precisos para medir el tiempo (reloj atómico) y el progreso en la tecnología espacial.

2.2.2 Reloj Atómico

En 1944, I. I. Rabi fue premiado con el Nobel de ese año por el desarrollo de la técnica de resonancia magnética para medir las frecuencias resonantes de los átomos. I. I. Rabi sugirió que debido a la precisión de las resonancias atómicas, estas podrían ser utilizadas para crear relojes extraordinariamente precisos.

En 1948 fue fabricado en EEUU, bajo la tutela de la Oficina Nacional de Normas, el primer reloj atómico. Utilizaba moléculas de amoniaco pero nunca fue usado ya que su precisión era inferior a la de un reloj de cuarzo común.

El primer reloj atómico práctico fue creado en 1955, en Gran Bretaña. Utilizaba unas frecuencias resonantes de cesio y tenía una precisión de un segundo en 300 años.

A partir de aquí nuevas versiones mejoradas de relojes atómicos comenzaron a surgir empleando frecuencias resonantes de cesio y rubidio entre otros componentes.

2.2.3 Desarrollo de la era espacial

El inicio de la navegación por satélite se produjo cuando los soviéticos lanzaron en 1957 el primer satélite en órbita terrestre. Su nombre, el Sputnik I.

No perdió mucho tiempo EEUU en ponerse en marcha y rápidamente encargó a varios investigadores su estudio.

No tardaron mucho en descubrir que conociendo la posición exacta de la tierra donde nos encontramos y midiendo, según se acercaba y se alejaba el Sputnik I, el desplazamiento de la señal de radio transmitida por el mismo podríamos conocer la posición exacta en la que se encontraba el satélite. Dicho desplazamiento en la frecuencia fue descubierto por Christian Doppler en 1842 y es conocido como el efecto Doppler.

En 1965 la US Navy (marina estadounidense) desarrolló el sistema TRANSIT como solución a la demanda de poseer un sistema de navegación fiable para submarinos que podían acometer inmersiones de larga duración. El tiempo empleado por el sistema para conocer la posición de los submarinos era de entre los 6 y 10 minutos y se podía determinar su posición en dos dimensiones con una precisión de 25 metros.

El siguiente sistema de Posicionamiento Global desarrollado por el Departamento de Defensa Estadounidense fue el Navstar, el cual depende de satélites que llevan relojes atómicos a bordo, (como evolución de un concepto desarrollado con anterioridad en un programa de la Marina llamado TIMEMATION), también consta de estaciones terrestres que controlan el sistema y receptores para el usuario que no dependen de relojes atómicos.

En 1978 fue lanzado el primer satélite GPS, siendo en 1989 puesta en servicio una segunda generación de satélites conocidos como los Satélites del Bloque II.

Fue en 1983, a consecuencia de que las fuerzas soviéticas derribaron un avión de pasajeros civiles coreanos que penetró en su espacio aéreo debido a errores de navegación, cuando el presidente Estadounidense, Ronald Reagan, declaró que el sistema GPS estaría disponible para usos civiles.

Pero la tecnología GPS no iba a ser entregada al 100%. Errores de cronometraje conocidos como “selective availability” (SA) fueron aplicados a las señales de GPS limitando así su posicionamiento en el uso civil.

Sin embargo, en 1991, durante la Guerra del Golfo, EEUU se vio obligado a eliminar temporalmente los errores de cronometraje para el uso civil. Fue debido a que el GPS se convirtió en una herramienta indispensable para el ejército Estadounidense que al no contar con suficientes receptores militares para todas sus

tropas, se vieron obligados a recurrir al uso de receptores civiles que debían de tener una precisión idéntica a los receptores militares.

Finalmente en el año 2000 los errores de cronometraje (SA) fueron eliminados permanentemente y de manera global, aunque el servicio militar de los EEUU aún puede introducirlos en las señales de extensiones geográficas limitadas.

Figura 3. Concepto artístico de un satélite GPS [2]

2.3 Posicionamiento indoor basado en redes inalámbricas

En la situación actual en la que vivimos, donde el desarrollo tecnológico avanza a pasos agigantados, son múltiples y de distinta índole los campos que experimentan cambios.

En concreto, el campo de las comunicaciones, con el desarrollo de las redes de sensores, PDA's, tabletas, Smartphones, etc… se ha colocado a la vanguardia del futuro tecnológico.

Esta serie de cambios propician la aparición de nuevos sistemas. Un ejemplo claro en la actualidad, son los sistemas de posicionamiento de interiores, que gozan de una gran popularidad. Este tipo de sistemas está sufriendo una gran demanda en los últimos años debido a su aplicación en múltiple campos como:

☑ La industria y el transporte.

Aplicada a la gestión de flotas, organización de almacenes e inventarios. Se consigue una producción y una supervisión más eficiente, mejorando la seguridad y reduciendo los errores en las entregas.

☑ Servicios sanitarios y de emergencia.

Seguimiento del personal sanitario que facilita la localización de los profesionales del sector. En los centros hospitalarios se aumenta el rendimiento en el cuidado del paciente y se mejora en la información sanitaria.

☑ Logística

En las áreas donde existe una gran variedad de productos, cuya movilidad es continua, dicha herramienta nos facilita la ubicación de artículos, equipos y personal.

☑ Domótica

Permite la activación automática de aplicaciones cuando el usuario se encuentra en un área determinada. Dicha herramienta aporta a los hogares rapidez y eficiencia en las acciones.

Hemos mencionado los campos donde su aplicación parece más intuitiva pero son innumerables los campos en los que su aplicación puede ser muy ventajosa.

Estás herramientas de posicionamiento en interiores basada en redes inalámbricas pueden hacer uso de distintos tipos de señal, pero la más comúnmente usada es la WIFI.

Su mecanismo de funcionamiento, mediante redes inalámbricas WIFI, se basa en la medición de las potencias de señal de los diferentes puntos de acceso ubicados en entorno que rodea al usuario.

Una gran ventaja que aporta el uso de la tecnología WIFI es que requiere poca infraestructura. Estas redes ofrecen facilidad de ampliación y se configuran de forma rápida y sencilla.

Además la red Wireless utiliza frecuencias libres (2,4 GHz) y no necesita licencia. La localización se puede realizar desde un dispositivo móvil que hace uso de dicha red Wireless para conectarse con el entorno.

Una de las desventajas de estas redes es que las interferencias y las fluctuaciones de potencia son habituales. Dichas interferencias son provocadas por elementos de todo tipo:

☑ Elementos fijos (tabiques, paredes,…)

☑ Elementos móviles (mesas, armarios, sofás,…)

☑ Personas en movimiento dentro y fuera de la estancia.

Es frecuente la presencia de dichos elementos en las estancias donde se propaga la señal WIFI y que producen una atenuación ocasionando el conocido como fenómeno multitrayecto, por el que las interferencias de una señal y sus ecos distorsionan la relación distancia-potencia.

Además solo los usuarios con capacidad de conexión a una red Wireless, podrán optar al uso de esta herramienta de posicionamiento en interiores.

El primer paso para crear una herramienta de este tipo es realizar una recopilación de potencias de señal que reciben el nombre de datos de entrenamiento. Dichos datos de entrenamiento, que deben de estar perfectamente asociados a las estancias donde fueron tomados, nos sirven para crear una base de datos.

La clasificación de muestras individuales, conocidas como muestras de test, se realiza mediante un algoritmo de clasificación multiclase basado en la base de datos, mencionada anteriormente, que almacena los datos de las estancias muestreadas y determina a qué estancia pertenece cada muestra individual.

Existen muchos algoritmos distintos para la clasificación multiclase que se pueden aplicar en este tipo de aplicaciones. Observando estudios teóricos, vemos que los algoritmos de clasificación más comúnmente utilizados son:

- ☑ Heurística de movimiento
- ☑ Heurística de proximidad
- ☑ Método de los vecinos más cercanos (k-vecinos)
- ☑ Redes Neuronales
- ☑ Maquinas de vector de soporte

Existen multitud de aplicaciones desarrolladas orientadas al posicionamiento en interiores. Cada una de ellas usa uno o varios algoritmos de clasificación que, en la mayoría de los casos, no se especifican. A continuación mostraremos las aplicaciones de posicionamiento de interiores que más renombre poseen en el mercado.

2.3.1 Aplicaciones de posicionamiento indoor basados en WIFI

2.3.1.1 Ekahau HeatMapper

Ekahau HeatMapper es una herramienta de software libre para el mapeo de cobertura rápida y fácil de las redes WIFI. Se presenta de manera gratuita para el usuario y con una interfaz muy sencilla.

Permite crear mapas de calor que muestran la cobertura que tiene una red WIFI en un lugar determinado. Ekahau HeatMapper™ utiliza su adaptador integrado de red inalámbrica, por lo tanto, todo lo que se necesita es un ordenador portátil que trabaje con Windows y soporte tecnología inalámbrica. Una vez instalado, es necesario realizar un entrenamiento que ayude a la aplicación a calibrar los mapas de las estancias.

Dicha aplicación soporta las normas 802.11n/g/a/b, es decir, los estándares de la IEEE más utilizados por lo que no tiene problemas de funcionamiento y compatibilidades.

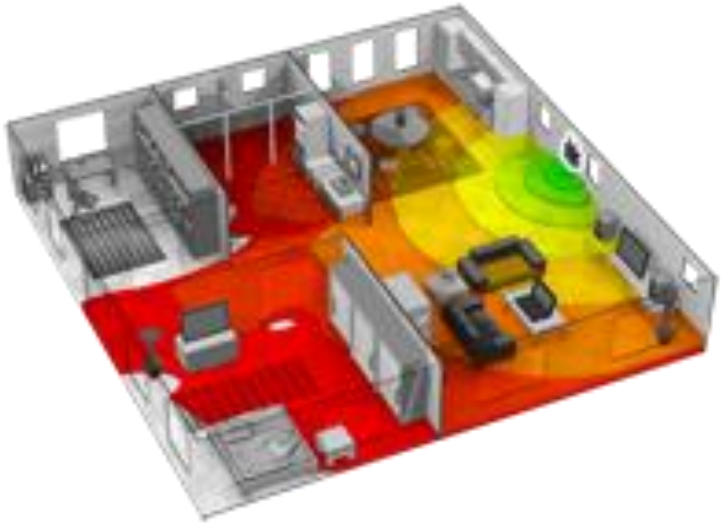


Figura 4. Ejemplo de un HEATMAPPER de Ekahau [6]

El sistema consta de tres elementos básicos:

☑ EPE (Ekahau Positioning Engine).

Su papel consiste en hacer de plataforma de localización. Desde aquí se calcula y se controla la posición de los clientes del sistema.

☑ Access Point.

Establece la conexión entre la red cableada y los clientes. La condición que impone el sistema es que al menos 3 puntos de acceso pertenezcan a nuestra red.

☑ Ekahau Client.

Software que adjunta una interfaz de red que incluye, para poder ser localizado, un transceptor radio y una antena.

Ekahau HeatMapper™ también proporciona una visión en tiempo real para todos los puntos de acceso y sus configuraciones. Permite la localización simultánea de múltiples dispositivos y sobre el mismo mapa de situación, ya sean dispositivos activos o pasivos.

Puede ser instalada en PDA's, Smartphone, portátiles,… pero con la restricción que el software Windows nos impone.

Uno de los aspectos más importantes de esta aplicación es que permite visualizar las configuraciones de seguridad de una red, ofreciendo la posibilidad de verificar el nivel de seguridad de la red a la que nos vamos a conectar.

Localiza los Access Point indicando y en cuál de ellos la potencia de señal es más fuerte.

Un resumen de lo que Ekahau HeatMapper es capaz de hacer es:

- Muestra cobertura WIFI en un mapa.
- Localiza todos los puntos de acceso.
- Detecta la configuración de seguridad y encuentra redes abiertas.
- Diseñado para el hogar y la pequeña oficina.
- Soporta 802.11n, así como a / b / g.

2.3.1.2 Gecko™.

Gecko™ es un sistema de posicionamiento indoor desarrollado por la empresa sueca Qubulus™. Es una aplicación Android que solo se encuentra disponible en la página web del desarrollador y no, como suele ser habitual, en punto de distribución estándar de aplicaciones Android, Google Play.

Figura 5. Logotipo de Qubulus [7]

Allí, previo registro, se puede descargar la aplicación disponible para los terminales Android. Se encuentra en dos versiones, la versión de prueba y la versión BETA. La versión de prueba es gratuita y permite realizar el muestreo de la superficie que se desee almacenando los datos en la aplicación pero sin tener acceso a ellos.

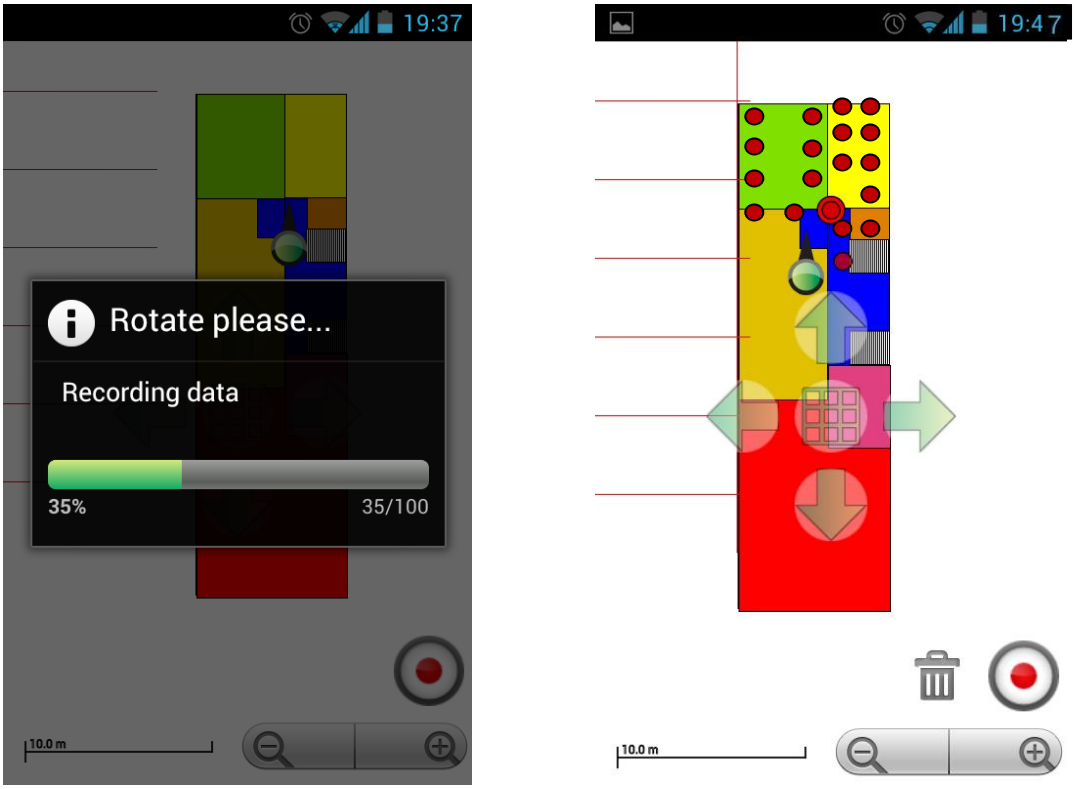
El uso de la aplicación es muy sencillo e intuitivo, pero la puesta en funcionamiento no lo es tanto. Lo primero que se debe hacer es crear un mapa a escala de la superficie que se vaya a muestrear. Posteriormente, usando Google Earth, se posicionará ese mapa de manera exacta sobre la superficie real que se vaya a muestrear. Se seleccionarán las coordenadas de la esquina superior izquierda y de la esquina inferior derecha que se introducirán en un archivo que a su vez estará en el interior de una carpeta junto con el mapa a escala de la superficie a muestrear.

Figura 6. Muestra del uso de Gecko [8]

Desde la página web se puede acceder a un canal Youtube que la empresa tiene habilitado para la explicación paso de paso de la puesta en funcionamiento que resulta bastante útil.

Una vez llegamos a este punto su funcionamiento es muy sencillo, simplemente indicaremos a la aplicación el punto del mapa donde nos encontramos y comenzamos a muestrear. La aplicación realiza la recopilación de potencias de señal WIFI existentes y las asocia al punto del mapa donde estamos muestreando la señal y que le hemos indicado con anterioridad. El proceso es el mismo para cada punto de la superficie, cuantos más puntos muestreemos mejores, resultados obtendremos.

Los datos obtenidos pueden ser exportados de manera sencilla por el usuario con la opción "Export data". Se generan archivos .dbe y .ebc, que son archivos encriptados con contraseña, por lo que su contenido no es accesible para el usuario.



(a) (b)

Figura 7. Recopilación de datos mediante el uso de Gecko [8]

El uso de la aplicación BETA es de pago y permite, una vez realizado el muestreo, comprobar su localización en el mapa mencionado con anterioridad utilizando la opción de "Location Mode", pero el precio de su licencia es totalmente abusivo para un particular. El código de la aplicación no se encuentra abierto, ni si quiera para versión BETA.

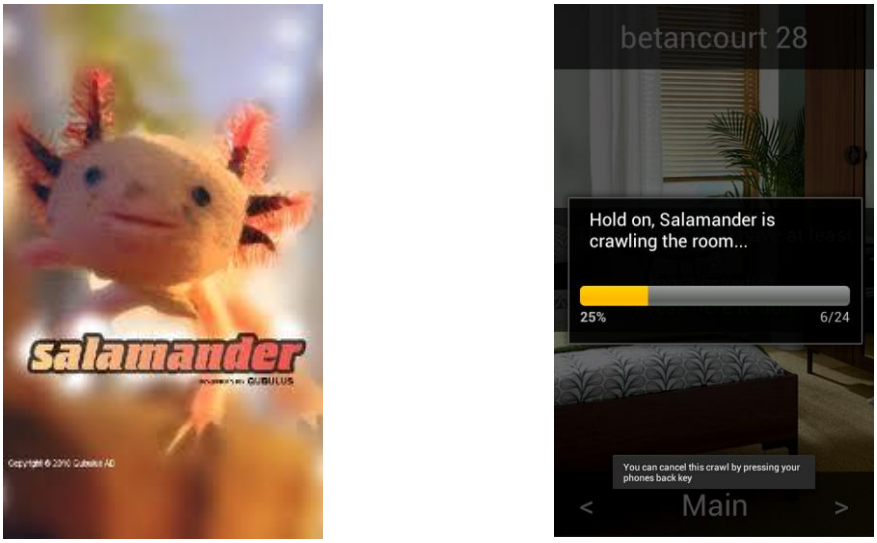
2.3.1.3 Salamander

Es otra aplicación de posicionamiento en interiores desarrollada por la misma empresa sueca que Gecko, es decir, Qubulus.

Figura 8. Logotipo de Salamander [7]

Dicha aplicación se encuentra disponible en el market de la plataforma Android, conocido como Google Play. Su uso es mucho más intuitivo y sencillo que el de su hermano mayor Gecko, pero la precisión que posee es mucho menor. Esta aplicación es capaz de diferenciar las distintas estancias que forman una superficie interior (baño, comedor, cocina).

Una vez se tiene instalada la aplicación, su uso es muy sencillo. Solamente se tiene que poner nombre a la estancia en la que nos se encuentra el usuario y realizar todas las medidas, dentro de la misma, que desee. La aplicación solicita un número mínimo de 3 muestreos para poder diferenciar estancias entre sí.

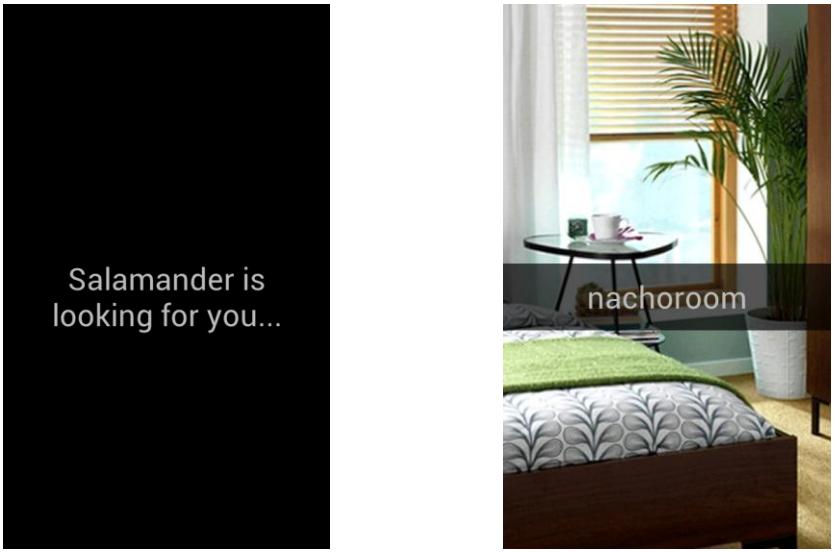


(a) **(b)**

Figura 9. Muestra del uso de Salamander [8]

Una vez recopilados los datos de entrenamiento, se almacenan en la aplicación, la cual nos ofrece la posibilidad de buscarnos a nosotros mismos y decir en que estancia nos encontramos. Su fiabilidad para distinguir estancias no muy distantes entre sí o que comparten unas intensidades de señal WIFI muy parecidas no es muy alta, cerca del 60% en estático e inferior si estamos en movimiento.

Sin embargo a pesar de su poca fiabilidad en lo que se refiere a la localización, es altamente útil para recopilar medidas para el entrenamiento de algoritmos, ya que con una sola petición, toma 24 medidas y almacena su promedio con lo que se obtienen resultados razonablemente estables.



(a) **(b)**

Figura 10. Muestra del uso de Salamander [8]

Los datos de potencia que recopilemos con la aplicación no están accesibles para el usuario de la misma. Sin embargo, existen otros métodos alternativos para su extracción.

```
paparoom,1,241940249752966,70.16666666666667,70686632816340,81.08333333333333,112391412317,85.66666666666667,117273617197156,88.41666666666667,
puertatorresquevedo,4,75465930928,70.0,73560807169,68.75,73560812945,66.41666666666667,220152338897888,73.75,73560812944,66.75,73560807168,68.83333333333333,118933703049,89.91666666666667,
puertatorresquevedo,4,75465930928,70.0,73560807169,71.08333333333333,73560812945,61.16666666666664,220152338897888,71.25,73560812944,61.333333333333336,73560807168,71.25,165791559601,89.41666666666667,
puertatorresquevedo,4,75465930928,70.0,73560807169,62.083333333333336,73560812945,73.58333333333333,220152338897888,68.08333333333333,73560812944,72.83333333333333,73560807168,62.083333333333336,
puertatorresquevedo,4,75465930928,70.0,73560807169,66.0,73560812945,66.83333333333333,220152338897888,72.33333333333333,73560812944,67.08333333333333,73560807168,65.91666666666667,
puertatorresquevedo,4,75465930928,70.0,73560807169,72.75,73560812945,66.75,220152338897888,61.833333333333336,73560812944,66.33333333333333,73560807168,72.58333333333333,12070791131,89.75,
puertatorresquevedo,4,75465930928,70.0,73560807169,69.41666666666667,73560812945,64.75,220152338897888,74.33333333333333,73560812944,64.58333333333333,73560807168,69.16666666666667,
```

Figura 11. Archivo de datos recopilados por Salamander [8]

Una vez recuperados los datos de potencia recopilados por Salamander™ nos encontramos el problema de su lectura, ya que se nos presenta en un formato poco agradable. Se creó un programa mediante Python que realizaba una lectura de los datos del archivo y presentaba dicha información en un formato más práctico.

Lo que conseguimos con esto es relacionar cada uno de los datos de potencia de señal WIFI con su SSID correspondiente o su MAC, además de con la estancia donde fueron tomadas. Más adelante explicaremos en qué consiste el SSID y la MAC de una red WIFI.

2.3.2 Aplicaciones de captura de potencia de señal

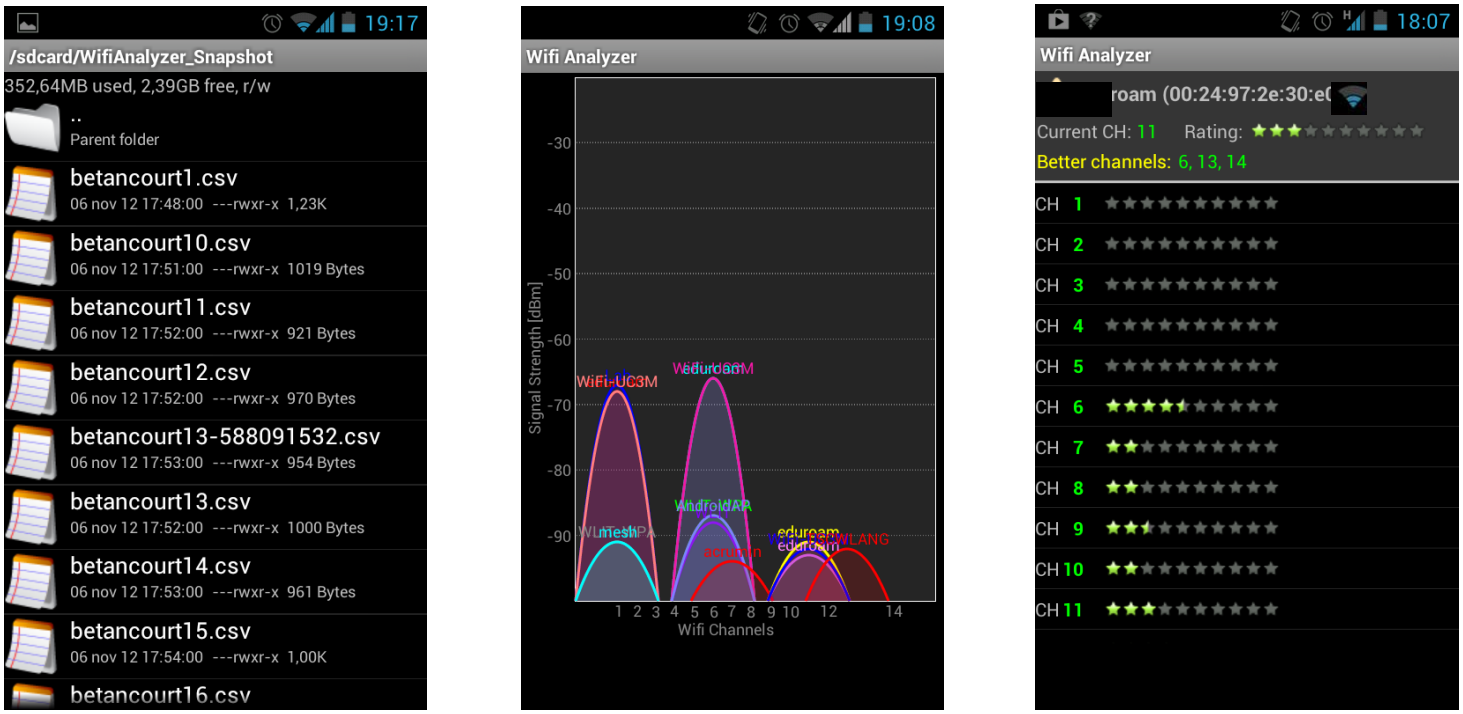
El entrenamiento de los algoritmos se realiza aportando una matriz de potencias recopiladas en las estancias donde queremos aplicar el posicionamiento indoor. Podemos usar aplicaciones propias de posicionamiento indoor y extraer los datos, como ya hemos visto en el apartado anterior, o podemos utilizar aplicaciones específicas de medición de potencias de señal WIFI.

Mostramos a continuación algunas de las aplicaciones existentes en el mercado para teléfonos móviles que usan la plataforma Android.

2.3.2.1 WIFI Analyzer

Es una aplicación que nos permite visualizar de manera gráfica las potencias de señal emitidas por los distintos puntos de acceso o router inalámbricos.

Además esta aplicación nos permite ver cuáles son los canales que deberían de usar los router inalámbricos para favorecer así la transmisión de la señal. También nos permite extraer toda esa información en archivos .csv, perfectos para realizar una matriz de potencias en Excel y posteriormente usarla en Matlab.



(a) (b) (c)

Figura 12.Muestra del uso de WIFI Analyzer. [8]

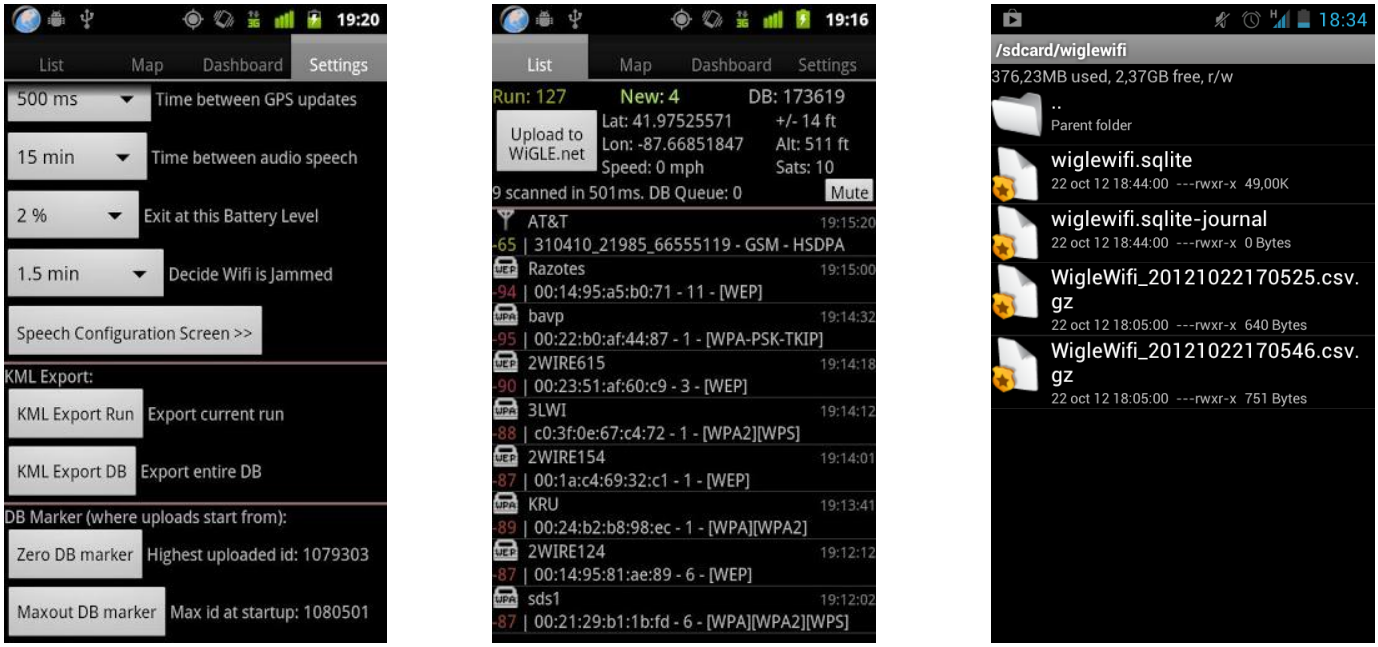
2.3.2.1 Wigle WIFI

Es otra de las aplicaciones que más éxito tiene en el mercado de aplicaciones de la plataforma Android.

Esta aplicación permite realizar barridos de señal al antojo del usuario proporcionando muchos datos de las redes WIFI que lo rodean. Además entrega datos de posicionamiento GPS del lugar donde nos encontramos. Dichos datos son accesibles para el usuario y se extraen en formato .csv.

El principal problema de esta aplicación es que requiere para su uso la disponibilidad de una señal GPS que adjunta a cada muestreo. La aplicación tarda varios minutos en completar la conexión GPS y posteriormente cualquier movimiento o desplazamiento hace que dicha conexión se caiga y la aplicación deje de funcionar.

Es una aplicación que en óptimas condiciones es muy interesante pero si queremos exprimir su uso para la creación de una herramienta de posicionamiento indoor, no parece la más adecuada.



(a) (b) (c)

Figura 13.Muestra del uso de WIFI Analyzer. [8]. y [9].

3. Planteamiento del problema: requisitos y restricciones

Nuestra principal meta es desarrollar una herramienta que, basándose en los datos reales de potencia de señal WIFI almacenados en matrices con las que se ha entrenado a una serie de algoritmos, sea capaz de averiguar la estancia donde se encuentra cualquier persona u objeto analizando la potencia de señal WIFI que su dispositivo móvil está recibiendo de los distintos puntos de acceso a la red Wireless.

Para este proyecto hemos de destacar que los datos que vamos a emplear para el entrenamiento de los algoritmos de clasificación van a ser reales y no simulados. Nuestra intención es la de ofrecer unos resultados acordes a la realidad y que tengan una credibilidad suficiente para que en caso de seguir adelante con una evolución en el desarrollo de la herramienta no existan dudas acerca de su fiabilidad.

La creación de dicha herramienta ha sido condicionada por una serie de requisitos y restricciones. Todos los requisitos han sido impuestos por nuestra parte, acordes a la forma en la vamos a desarrollar la creación de esta herramienta y con el fin de cumplir el objetivo marcado, pero algunas de las restricciones nos han sido impuestas por las características del servicio GPS, la tecnología WIFI o las aplicaciones empleadas para la recopilación de los datos.

Una vez comentadas con anterioridad las características de las aplicaciones para la recopilación de datos existentes en el mercado, debemos poner en contexto cual es el funcionamiento y las características que poseen tanto el GPS como una red inalámbrica WIFI para que se comprendan el resto de restricciones con las que cuenta nuestra herramienta.

3.1 Funcionamiento del GPS y Aplicaciones del GPS

Los satélites GPS puestos en órbita transmiten su posición exacta y una señal de tiempo extremadamente precisa (usan relojes atómicos) que llega a los receptores situados en la superficie terrestre.

Una vez se posee dicha información los receptores GPS pueden calcular su distancia al satélite, y combinando la misma información de 4 satélites distintos, el receptor puede calcular su posición exacta en la tierra usando el proceso de trilateración.

La trilateración se basa en el concepto de que si un receptor conoce la distancia a un satélite, sabe que su posición se encuentra sobre una esfera con centro en el satélite y de radio igual a la distancia al mismo.

Si obtenemos la misma información de un segundo satélite, podemos mejorar la precisión de nuestra posición viendo el área que tienen en común las dos esferas.

Al añadir el tercer satélite reducimos nuestra posición exacta a dos posibles puntos. Esos puntos son en los que las tres esferas creadas se cruzan.

La cuarta medida se usa para saber cuál de los puntos es el que nos da la ubicación exacta, ya que esa cuarta esfera solo cortaría la otras tres en uno de esos dos puntos, pero generalmente uno de ellos representa una posición absurda y se auto-descarta. Aun así la cuarta medida se sigue considerando necesaria ya que en el caso de un error de medición, si contamos con las cuatro medidas y nos apoyamos en el ordenador del receptor, aplicando las correcciones necesarias para obtener la posición correcta, este hará que las cuatro esferas se corten en un único punto obteniendo la ubicación exacta. La cuarta medida también da a conocer la hora.

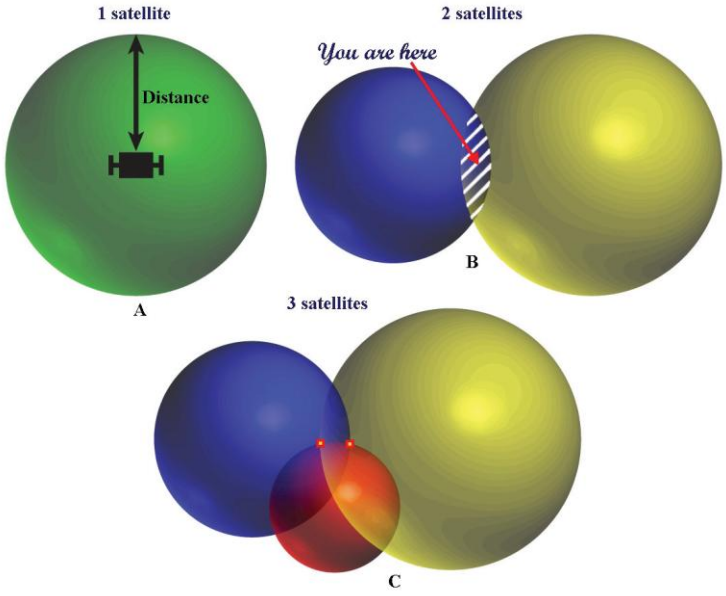


Figura 14. Cálculo de la posición usando 3 satélites. [4]

Otro de los errores que se pueden dar son los provocados por la interferencia atmosférica y reflejos de obstáculos en la tierra como árboles y edificios. Utilizando métodos como la corrección por frecuencia doble o filtros de señales podemos disminuir sus efectos.

Una de las características que se está desarrollando es el aumento de la precisión para conseguir unos sistemas GPS más precisos y fiables. Podemos destacar el GPS diferencial (DGPS) y el Sistema de Aumento de Áreas Amplias (WASS).

Los campos donde se puede aplicar es GPS son muy variados:

☑ **Agricultura**

Aumenta la producción y ayuda a mejorar la eficiencia de los métodos de cultivo. Este tipo de sistemas, ubicados en sofisticados sistemas de información geográfica, son capaces de recoger datos sobre condiciones del suelo, humedad, temperatura y muchas más variables.

☑ **Navegación en tierra y mar**

Aparte de los usos básicos para determinar la posición, en este campo podemos ayudarnos del GPS para mejorar las cartas marinas, marcar objetos sumergidos como obstáculos y guiar sistemas de piloto automático. Además es extremadamente útil para los equipos del salvamento marítimo que reducen el tiempo de llegada al punto de donde ha surgido el conflicto.

☑ **Usos militares**

Como no podía ser de otra manera, ya que fueron ellos los precursores, el GPS tiene múltiples funcionalidades en el ámbito militar. Aparte de la más común como la navegación general, el GPS también es usado para dirigir proyectiles o bombas inteligentes hacia los destinatarios y despliegue de ejércitos en campos de batalla.

☑ **Ciencias**

Este es otro campo donde sus aplicaciones son innumerables. Es usado desde para construir mapas, localizar estaciones de muestreo, definir límites geográficos hasta para seguir poblaciones animales, realizar análisis espaciales de rasgos naturales o aplicaciones en el campo de la sismología.

3.2 Funcionamiento, arquitectura y componentes de una red WIFI

Las WLAN quedan encuadradas dentro de los estándares desarrollados por la IEEE. Dichos estándares referidos a las WLAN son los 802.11 X, cuyas características están explicadas en el estado del arte. Posteriormente varios organismos han desarrollado más estándares que versionan los creados por IEEE.

Una WLAN tiene varias configuraciones posibles pero existen 3 que destacan por encima del resto.

☑ **AD-HOC**

Los terminales se comunican libremente entre sí y suele encontrarse en entorno militares, operaciones de emergencia, comunicación entre vehículos.

☑ **Infraestructura**

Aquí los equipos están conectados a uno o más puntos de acceso normalmente conectados a una red cableada encargada del control de acceso al medio. Suele encontrarse en hogares, empresas o instituciones públicas.

☑ **Entre varias WLAN**

Se interconectan LAN's o WLAN's distantes.

Todavía tenemos que presentar algunos actores más que se encuentran presentes en una red WIFI y que son necesarios para el correcto funcionamiento de la misma.

☑ **Los adaptadores inalámbricos** (Network interface controler) ejercen el control de la interfaz de red. Son tarjetas de red que se ajustan a la normativa del estándar 802.11 y que permite a un equipo establecer una conexión con una red inalámbrica. Existen diversidad de formatos de presentación de dichos adaptadores (tarjetas PCI, USB,…..)

☑ **Puntos de acceso** (Access Point): permiten a los equipos acceder a una red, significando el nexo de unión entre el usuario y la red.

☑ **MAC** (media Access control) es un identificador de 48 bits de longitud dividido en 6 bloques hexadecimales que corresponde, de forma única, a una tarjeta o dispositivo de red. Estas direcciones son únicas a nivel mundial y son escritas en binario en el instante de fabricación del hardware.

☑ **SSID** (Service Set Identifier) es un identificador añadido a la red WIFI que es customizable a gusto del usuario. Todas la redes WIFI traen por defecto una SSID que podemos cambiar para hacerla más familiar.

Un punto muy importante a tener en cuenta es que el protocolo 802.11 permite en su capa MAC comprobar la integridad de los datos transferidos. Es muy frecuente el envío de paquetes de gran tamaño por redes inalámbricas, lo que provoca que el índice de errores de transmisión incremente.

Para evitar perder paquetes enteros existe un mecanismo de fragmentación que divide la trama en fragmentos.

Cada trama se subdivide en tres partes:

☑ Un encabezado de 30 bytes

☑ Un cuerpo donde se transmite el mensaje/datos

☑ Una secuencia de verificación para corregir los errores.

3.3 Requisitos y restricciones

Una vez explicado el funcionamiento general del GPS y de una red WIFI procederemos a exponer la lista de los requisitos y restricciones generales de nuestro proyecto.

1. Los datos de potencia de señal WIFI recopilados han de ser reales.
2. La recopilación de las medidas de potencia de señal se realizarán con un teléfono móvil cuyo sistema operativo sea Android.
3. Se necesitarán aplicaciones compatibles con dicho sistema operativo que nos permitan la medición y extracción de los datos de potencia.
4. Todos los datos de entrenamiento recopilados han de estar perfectamente etiquetados con la estancia donde fueron tomados y perfectamente relacionados con su SSID correspondiente o, en su defecto, su MAC correspondiente.
5. Las estancias seleccionadas han de tener diversidad en el número de puntos de acceso, es decir, utilizaremos estancias muy densamente pobladas de distintas señales WIFI y no tan pobladas. El objetivo es asegurarnos de que la herramienta funcionará en cualquier estancia independientemente de la densidad de señal existente
6. Es necesario el uso de un mínimo de dos aplicaciones para recopilar dos datos de entrenamiento en las estancias conseguir así una alta fiabilidad en los resultados obtenidos.
7. Será necesaria realizar el entrenamiento y la posterior clasificación de los datos recopilados, por ambas aplicaciones, con al menos 4 algoritmos de clasificación multiclase.
8. Los algoritmos de clasificación han de ser, aparte de eficaces, rápidos ya que el futuro de esta herramienta es estar instalada en un dispositivo móvil donde el usuario necesita la información al instante.
9. Para la localización del usuario, nos apoyaremos en los datos de entrenamiento almacenados con los que hemos entrenado a los algoritmos. Los distintos algoritmos determinarán la estancia de donde provienen el usuario analizando los vectores de potencias individuales de cada uno.
10. Los vectores de potencia de señal recopilados para el entrenamiento de los algoritmos quedarán almacenados en matrices de potencia medidas en dBm´s.

11. Los elementos que componen cada vector de potencias, tanto los destinados para entrenamiento como los destinados a clasificar, han de encontrarse por encima de los -90dBm, ya que las aplicaciones no detectan señales por debajo de estos valores.

12. La herramienta solo podrá localizar, de manera fiable, a los usuarios de la misma si estos se encuentran en las zonas definidas para la captación de datos de entrenamiento.

13. No buscamos coordenadas, como nos da el GPS, buscamos que el algoritmo nos facilite la estancia donde nos encontramos.

4 Diseño y resultados de la solución técnica

4.1 Diseño de la herramienta

Observando el estado del arte de las herramientas de posicionamiento indoor, la intención de este proyecto es el desarrollo de una herramienta que ofrezca un servicio similar que dichas aplicaciones, disponibles en el mercado, pero intentando mejorar la calidad de las mismas.

Cada herramienta de posicionamiento indoor tiene unas características particulares que la diferencian de las demás. Pero la principal diferencia radica en los algoritmos de entrenamiento y clasificación de potencia de señal que usan.

Son pocas o ninguna las aplicaciones que hacen público el algoritmo en el que se basan debido a su importancia ya que dichos algoritmos afectan, en gran parte, a los puntos críticos como la precisión de la herramienta y la velocidad de respuesta de la misma indicando al usuario su posición.

Con el objetivo ya mencionado, hemos decidido emplear cuatro algoritmos de clasificación multiclase para ser entrenados con muestras reales tomadas por nosotros y recopiladas mediante dos aplicaciones:

- ☑ Salamander
- ☑ WIFI Analyzer

Los algoritmos de clasificación multiclase empleados van desde métodos clásicos como el método de los vecinos más cercanos (k-vecinos) hasta lo que podríamos llamar el estado del arte, en cuanto a clasificación multiclase se refiere, las maquinas de vectores de soporte (SVM).

Los otros dos algoritmos empleados son la regresión multinomial y el algoritmo Naive Bayes.

Las muestras tomadas han sido recopiladas procedentes de distinto tipo de estancias. La superficie de dichas estancias va desde los 10,60m^2 de una habitación de una vivienda unifamiliar, hasta el hall de los Edificios Betancourt y Torres Quevedo de la Universidad Carlos III de Madrid.

El objetivo es que nuestra herramienta, dado un vector de potencias individual y siendo este analizado por los algoritmos (ya entrenados), nos indique la estancia en la que dicha muestra fue tomada. No buscamos coordenadas de posición X e Y, sino lograr diferenciar las estancias.

4.2 Explicación de los algoritmos de clasificación multiclase utilizados

4.2.1 KNN

La idea básica de este algoritmo es que una muestra nueva se va a clasificar en la clase más frecuente, a la que pertenecen sus K vecinos más cercanos.

Una de sus principales atracciones de este algoritmo de clasificación multiclase es su simplicidad y potencia.

Este algoritmo tiene una funcionalidad que se explica en los siguientes pasos:

- Disponemos de un conjunto de datos de entrenamiento o muestras ya clasificadas. En nuestro caso, la matriz de intensidades de señal WIFI.
- Se calcula la distancia de la muestra a clasificar (conocida como la muestra de test), a las muestras del conjunto de datos de entrenamiento.
- Se calculará la distancia (Matlab nos proporciona la euclídea por defecto, pero ofrece modalidades) de la nueva muestra (test) respecto a cada una de todas las muestras con las que ya cuenta el algoritmo.
- Aquí es donde entra en acción el valor K que introduce el algoritmo. La clase resultante de la muestra a clasificar (la clase que se asignará a la muestra de test) será la calculada respecto a las K instancias que minimicen la distancia o K vecinos más cercanos. La clase finalmente elegida para esa muestra de test será la que más veces se repita en los K vecinos más cercanos

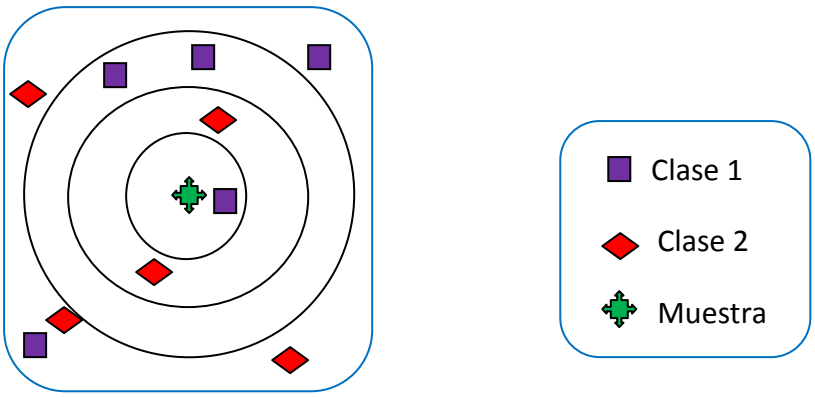


Figura 15. Ejemplo de algoritmo KNN. [8]

Si utilizamos un K=1, es decir, solo tener en cuenta el círculo más pequeño la clase que le asignaríamos a la muestra de test sería la clase 1, ya que es la clase de su vecino más cercano.

Por el contrario si eligiésemos K=3, la clase que le asignaríamos, teniendo en cuenta los tres mejores vecinos, sería la 2 debido a que dos de ellos son de dicha clase.

Si pasásemos al siguiente nivel usando un K=5 de nuevo la clase que se le asignaría a las muestra de test sería la 1.

De esta manera podemos observar como malas elecciones del valor de K puede conducirnos a resultados erróneos. La elección de dicho valor depende fundamentalmente de los datos. Usar valores grandes de K reduce el efecto ruido en la clasificación pero introduce una gran desventaja al provocar la aparición de límites entre clases parecidas, es decir, al tener en cuenta tantos vecinos para la toma de la decisión puede darse el caso de que exista un empate entre vecinos de dos clases.

4.2.1.1 Distancias KNN

Todo algoritmo de clasificación precisa de una métrica que nos permita comparar las distancias existentes entre los objetos. Existe un amplio abanico de posibilidades para determinar la distancia a usar, la más común es la distancia euclídea no ponderada aunque existen alternativas que podemos comprobar con el objetivo de ver si se produce una mejora en la clasificación de los datos.

- ☑ **Distancia euclídea no ponderada.**

Es la más utilizada en el ámbito del cálculo de distancias al ser la más cómoda. Se define como la longitud del segmento que une dos puntos y su función se expresa de la siguiente manera:

$$d(A,B) \equiv \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2} = \sqrt{(A-B)^T (A-B)}$$

Figura 16. Fórmula de la distancia Euclídea no ponderada. [11]

Otras distancias que dicho algoritmo nos ofrece para su cálculo en Matlab son:

- ☑ Cityblock
- ☑ Cosine
- ☑ Correlation

4.2.2 Naive Bayes.

Cuando nos referimos a un clasificador Naive Bayes estamos hablando de un simple clasificador probabilístico que se basa en la aplicación del teorema de Bayes asumiendo la independencia de supuestos. También podemos hacer referencia a él como: "modelo de características independientes".

El clasificador Naive Bayes, dada una variable de clase, asume que la ausencia o presencia de una característica determinada no está relacionada con la ausencia o presencia de cualquier otra característica.

Por ejemplo, nos situamos en el caso de conocer las características que definen que un automóvil puede pasar a considerarse un todoterreno (ruedas miden X centímetros de anchura y perfil, cuando las dimensiones de la longitud del coche sobrepasan unas establecidas, cuando tiene determinado número de marchas, etc). El algoritmo Naive Bayes considera dichas características de manera independiente para la probabilidad de que dicho automóvil obtenga la calificación de todoterreno aunque alguna de ellas estén o no relacionadas entre sí.

Este clasificador puede ser entrenado de manera eficiente mediante un aprendizaje supervisado.

Las características más reseñables de los métodos bayesianos son:

☑ Cada ejemplo observado va a modificar la probabilidad (aumentándola o disminuyéndola) de que la hipótesis formulada sea correcta.

☑ Son métodos robustos frente al posible ruido presentes en los ejemplos de entrenamiento.

☑ Durante la valoración de cada hipótesis, permiten tener en cuenta el conocimiento a priori.

4.2.2.1 Clasificación de patrones en Naive Bayes.

De manera abstracta, podemos hablar de que el modelo de probabilidad para un clasificador es un modelo condicional del tipo:

$$p\,(C|\, f_1, f_2, f_3, f_4, \ldots, f_n)$$

Figura 17. Modelo Condicional para el clasificador Naive Bayes. [19]

El problema aparece cuando el índice n es demasiado grande o cuando una característica puede tomar demasiados valores distintos. Dadas estas circunstancias, para un modelo basado en tablas de probabilidad, su uso no es factible.

Por esta razón, se realiza una reformulación del modelo para ser más manejable. Aplicando el teorema de Bayes:

$$p(C|f_1, f_2, f_3, f_4, \ldots, f_n) = \frac{p(C)\, p(f_1, f_2, f_3, f_4, \ldots, f_n|C)}{p(f_1, f_2, f_3, f_4, \ldots, f_n)}$$

Figura 18. Fórmula resultante de aplicar el teorema de Bayes en el algoritmo explicado. [19]

Una vez observamos esta última fórmula vemos que el denominador no depende de C, y está formado por unos valores de las características (f) que ya conocemos. Por lo que debemos de mostrar atención en el numerador.

Un desarrollo más incisivo del numerador, teniendo en cuanta la independencia entre muestras, nos llevaría a:

$$p(C|f_1, f_2, f_3, f_4, \ldots, f_n) = p(C)\, p(f_1|C)\, p(f_2|C)\, p(f_3|C) \ldots\ldots\, p(f_n|C)$$

A partir de aquí podemos estimar $p(C)$ realizando la suma de las veces que aparece C en el conjunto de datos de entrenamiento y dividiéndolo por el número total de de ejemplos que forman este conjunto.

Si lo que queremos estimar es $p(C|f_1, f_2, f_3, f_4, \ldots, f_n)$ para conocer las veces en que para cada categoría aparecen los valores f, debo recorrer todos los datos de entrenamiento.

4.2.3 Regresión logística multinomial.

La tradición nos muestra que el modelado de variables dependientes politómicas, aquellas que cuentan con más de dos categorías, se viene realizando mediante el uso de análisis discriminantes. Pero en la actualidad, gracias al desarrollo experimentado en las técnicas de cálculo, el uso de modelos de regresión logística multinomial comienza a ser habitual.

Utilizado en modelos con variable dependiente de tipo nominal politómica, este nuevo ejemplo de algoritmo de clasificación multiclase está considerado una extensión multivariante de la regresión lógica binaria clásica.

Si se trata de variables independientes, estas pueden ser tanto continuas como categóricas.

El análisis de estos modelos se realiza eligiendo una categoría de referencia de la variable dependiente. Seguidamente se modelan varias ecuaciones de manera simultánea, una para cada una de las categorías restantes respecto a la de referencia.

A continuación, mostramos las etapas y requisitos de la regresión lógica multinomial:

- Recodificar las variables independientes categóricas en variables simuladas y la variable dependiente.
- Evaluar los efectos de confusión y la interacción del modelo explicativo.
- Analizar la fuerza, sentido y significación de los coeficientes, sus exponentes y estadísticos de prueba.

En nuestro caso no utilizamos todas las etapas de la regresión logística.

4.2.3.1 Formulación del método.

Para explicar el algoritmo vamos a servirnos de un ejemplo.

Tenemos 5 estancias distintas (clases) con sus correspondientes medidas a clasificar:

Figura 19a. Ejemplo de formulación del método. [13]

Considerando el comedor la categoría de referencia, se modelan varias ecuaciones de manera simultánea una para cada una de las categorías restantes respecto a la de referencia, en este caso 4.

Estas ecuaciones reciben el nombre de transformaciones logarítmicas (logits).

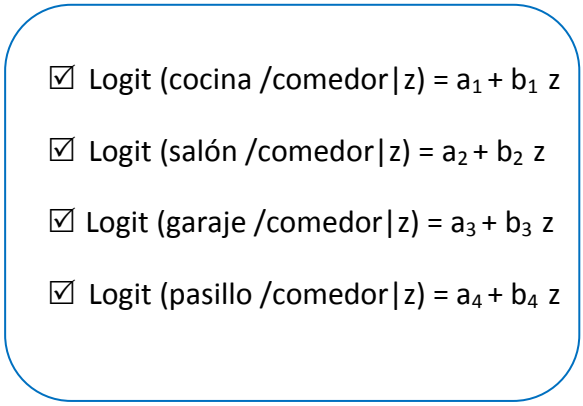


Figura 19b. Ejemplo de formulación del método. [13]

Cada logit tiene en común la covariable z, una constante a y un coeficiente b. Tanto a como b son diferentes en cada logit.

4.2.4 Máquinas de vectores de soporte.

Se refieren a un conjunto de algoritmos de aprendizaje supervisado y son métodos relacionados con la clasificación y la regresión.

Basándonos en una serie de datos de entrenamiento y etiquetándolos podemos entrenar una SVM para construir un modelo de predicción de clases. Una SVM construye un hiperplano o un conjunto de hiperplanos en un espacio de dimensionalidad muy alta. Si la separación entre las clases es buena, la clasificación será exitosa.

El SVM clásico solo funciona de forma binaria, es decir, tan solo clasifica dos clases. En nuestro caso tendremos un número de estancias muy superior a 2 y por lo tanto necesitamos una variación del algoritmo. Esa variante se llama SVM Multiclass.

Existen dos posibles variantes: One vs all o One vs One. Hemos utilizado la primera, One vs all. Este clasificador consiste en crear tantos algoritmos SVM como clases tenemos a clasificar.

Cada SVM creado para cada clase divide todas las muestras de entrenamiento en dos clases (independientemente del número de clases existentes en dicho). Es decir, trabaja como una SVM binaria que sólo distingue entre las muestras de su clase (por ejemplo, clase 1) y agrupa el resto de las muestras del conjunto en una misma clase (por ejemplo, clase 0) aunque pertenezcan a clases diferentes.

De esta manera somos capaces de diferenciar una clase respecto a todas, y al hacerlo tantas veces como clases tenemos conseguimos clasificar todas las muestras.

Este método de clasificación multiclase ha superado a las conocidas como Redes Neuronales y se sitúa como el estado del arte en cuanto a la clasificación multiclase se refiere.

4.3 Recopilación de medidas y resultados obtenidos.

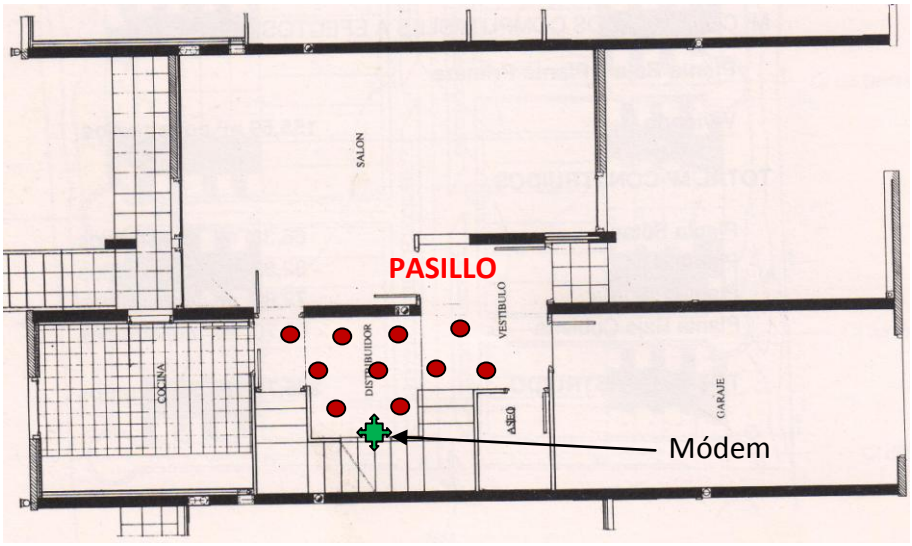
Las primeras medidas tomadas se realizaron en diferentes estancias de una vivienda unifamiliar. Fueron realizadas en tres plantas distintas:

1º Planta → Pasillo de 12 m^2

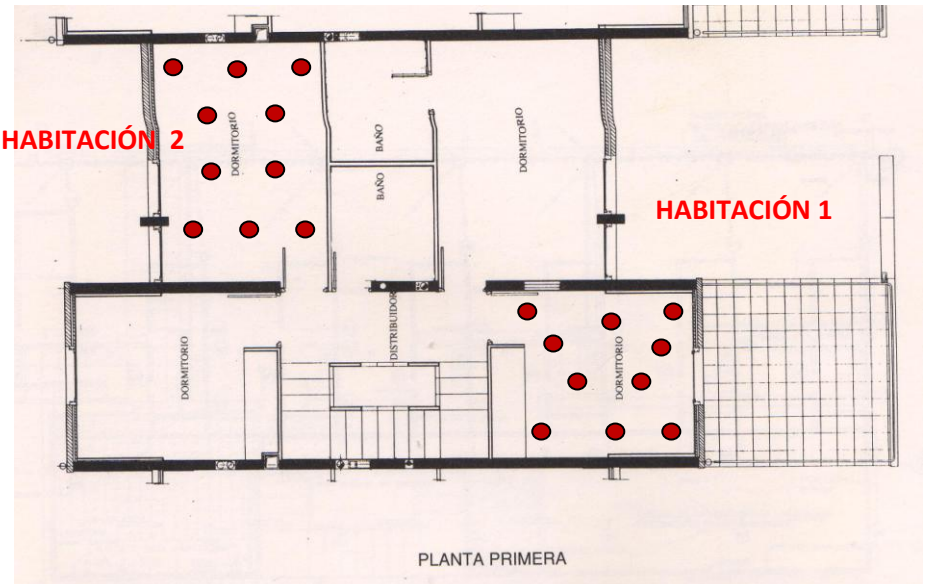
2º Planta → Dos habitaciones (dormitorios) de 10,60 m^2 y 12,60 m^2.

3º Planta → Buhardilla de 45 m^2

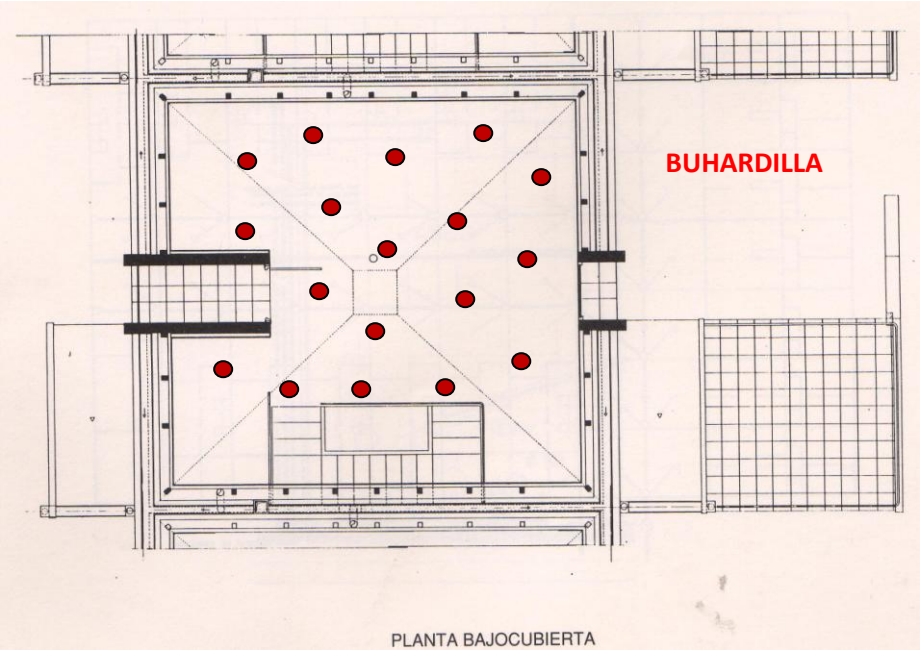
La recopilación de las medidas se realizó, como ya hemos comentado con anterioridad, con las aplicaciones Salamander y WIFI Analyzer. Se tomaron 10 muestras por estancia para el pasillo y las dos habitaciones, y 18 muestras en la Buhardilla.



(a)



(b)



(c)

Figura 20. Puntos de recopilación de potencias de señal WIFI en la vivienda unifamiliar. [8]

En los planos originales de la vivienda unifamiliar podemos observar los puntos de muestreo de cada estancia, los cuales, fueron los mismos durante la recopilación de datos mediante el empleo tanto de Salamander, como de WIFI Analyzer. De esta manera, posteriormente pudimos comparar los resultados evitando, en la medida de lo posible, haber tomado datos no reales, es decir, picos de señal de procedentes de distintos puntos de acceso que pudieran no servirnos.

Además los puntos de muestreo se tomaron de manera equidistante (1 metro / 1,5 metros) dependiendo de cada estancia, intentando, en la medida de lo posible abarcar la superficie completa.

Una vez teníamos en posesión las medidas, debíamos de exportar al ordenador dichos datos que las aplicaciones almacenan en la SD Card del teléfono móvil. Las exportaciones se realizaron en diferentes tipos de archivos, dependiendo de las facilidades que la aplicación empleada ofreciese.

Cada muestreo realizado era clasificado en un vector de potencias, asociando la potencia de señal (medida en dBm) a su identificador MAC en hexadecimal o al nombre del SSID con el que estuviese configurado. Por último cada uno de estos vectores de potencia quedaba etiquetado con la estancia donde fueron tomados.

La unión ordenada de todos estos vectores proporcionó una matriz de potencias de señal que usaríamos para el entrenamiento de los algoritmos de clasificación.

Para evaluar las prestaciones, utilizamos el método leave-one-out, con el que conseguimos que la muestra a clasificar no forme parte de los datos de entrenamiento del algoritmo usado. Es decir, si tenemos 100 muestras, para clasificar la primera de ellas entrenamos al algoritmo con las muestras comprendidas entre la 2 y la 100, dejando fuera del entrenamiento del algoritmo, que empleamos, la muestra a clasificar.

Este proceso se realiza para que el algoritmo no tenga conocimiento de la muestra que vamos a clasificar, porque si no, el acierto que nos mostraría el algoritmo estaría trucado indicando un acierto por encima del real. Además, de esta manera no es necesario que tomemos medidas de entrenamiento y medidas de clasificación por separado, sino que las mismas medidas nos valen para los dos ámbitos.

Las dimensiones totales de las matrices con la que entrenamos los algoritmos eran de 48 x 15 para WIFI Analyzer y de 48 x 7 para Salamander. Esta diferencia de medidas es debida a que la aplicación WIFI Analyzer detecta mayor número de puntos de acceso que la aplicación Salamander.

La descomposición de las matrices, dependiendo de la aplicación, es la siguiente:

- Para WIFI Analyzer:
  - ☑ 48 muestras (vectores fila).
  - ☑ 14 columnas (MAC´s).
  - ☑ 1 Columna final (Etiquetado de muestras).
- Para Salamander:
  - ☑ 48 muestras (vectores fila).
  - ☑ 6 columnas (MAC´s).
  - ☑ 1 Columna final (Etiquetado de muestras).

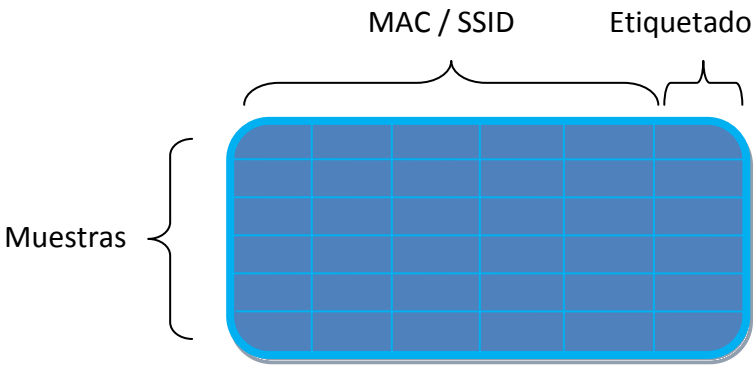


Figura 21. Tabla ilustrativa de organización de las muestras recopiladas. [8]

Lógicamente no en todas las muestras (vectores fila) se detectaba señal de todas las MAC's presentes en la tabla (dependiendo de la aplicación usada), sino que en la tabla se representan las distintas MAC´s registradas a lo largo de la recopilación de muestras por la vivienda y que dependiendo de la estancia donde estuviésemos detectaba señal o no. Aquellos lugares donde no se detectaba señal se ha representado con un valor nulo en la matriz de recopilación de datos.

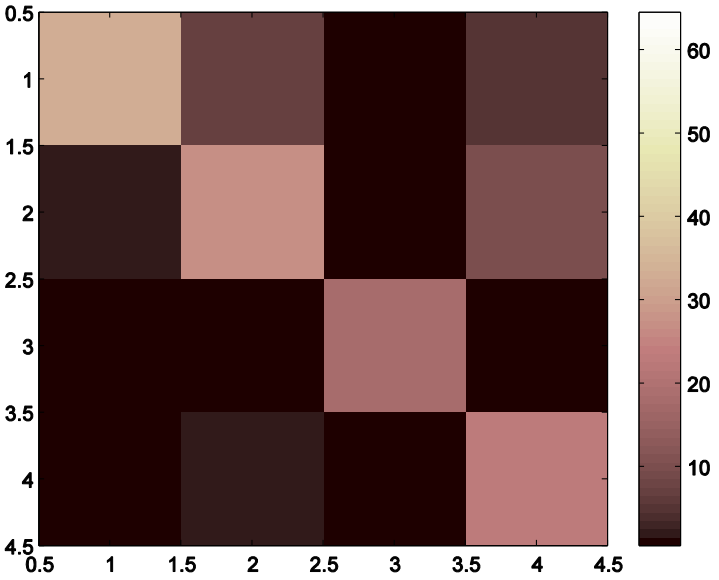
Los resultados obtenidos para cada algoritmo fueron los siguientes:

Algoritmos de clasificación	Datos de Salamander	Datos de WIFI Analyzer
KNN (K- Vecinos)	0,79	0,83
Naive Bayes	0,38	0,70
Multiclass SVM	0,60	0,75
Regresión multinomial	0,53	0,64

Figura 22. Tabla de acierto en clasificación de los algoritmos en una vivienda Unifamiliar. [8]

Las tasas de acierto en clasificación de los distintos algoritmos varían considerablemente, pero la mayoría de ellos son bastante aceptables teniendo en cuenta que la tasa de acierto aleatoria es 1/4.

Mediante unas matrices de confusión, correspondientes a los resultados ofrecidos por el algoritmo KNN para ambas aplicaciones, vamos a observar cuales son las muestras que más dudas generan a la hora de clasificarlas.

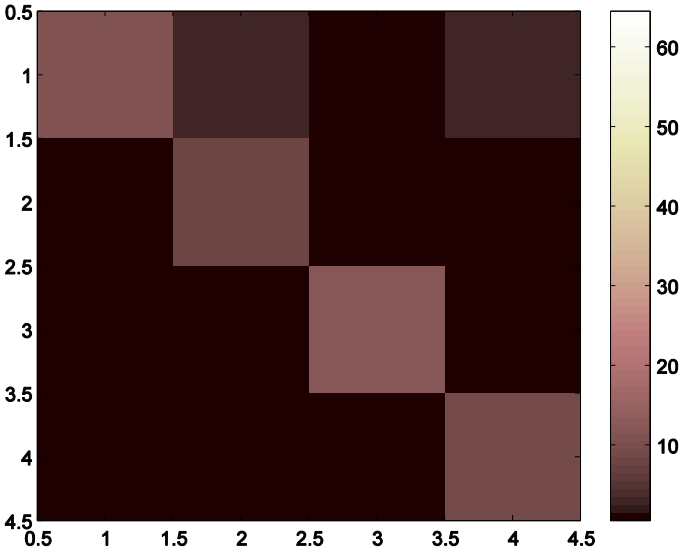


Salamander

Clase 1.....Habitación 1
Clase 2.....Buhardilla
Clase 3.....Pasillo
Clase 4.....Habitación 2

Podemos observar como las muestras que menor error de clasificación tienen son las recogidas en el pasillo, el resto de clases muestran errores de clasificación bastante apreciables. La presencia del módem en el pasillo es el principal motivo de tal diferencia de error entre clases.

Figura 23. Mapa de confusión del algoritmo KNN con datos recogidos mediante la aplicación Salamander. [8]



WIFI Analyzer

Clase 1…….Habitación 1
Clase 2…….Buhardilla
Clase 3…….Pasillo
Clase 4…….Habitación 2

Las muestras tomadas con esta aplicación parecen que generan algo menos de dudas que las tomadas con Salamander. Obervamos que se han corregido considerablemente los errores de clasificación para las muestras tomadas en las estancias 1,2 y 4 y que la 3 se mantiene bastante fiable,

Figura 24. Mapa de confusión del algoritmo KNN con datos recogidos mediante la aplicación WIFI Analyzer. [8]

Hemos mostrado las matrices de confusión del algoritmo KNN debido a que es el que mejores tasas de acierto, promediando las dos aplicaciones, ofrece. Además, en las matrices de confusión con tasas de acierto más bajas apenas se diferenciaban los errores de clasificación al ser prácticamente de un color homogéneo en su totalidad.

Una vez vimos los porcentajes de acierto obtenidos de los diferentes algoritmos clasificando estas primeras muestras, decidimos ampliar el número de estancias a clasificar y, a ser posible, el número de puntos de acceso que en dichas estancias se encontrasen.

De esta manera comprobaríamos si nuestra herramienta obtendría mejores resultados trabajando en emplazamientos con un mayor número de puntos de acceso a la red o por el contrario, sus porcentajes de aciertos se desplomarían. Decidimos trasladarnos al campus de la Universidad Carlos III de Madrid.

Hall Torres Quevedo
Cafetería Puerta 1
Cafetería Puerta 2
Hall Agustín de Betancourt
Reprografía

Figura 25. Plano de los puntos de recopilación de datos en el Campus de la Universidad. [8] y [14].

Los puntos que se muestran en la leyenda a la derecha de la imagen del campus de la Universidad fueron las estancias elegidas para la recopilación de nuevas potencias de señal.

Para esta segunda tanda de recopilación de datos tomamos entre 16 y 18 medidas en cada estancia intentando que las muestras fueran tomadas de manera equidistante (1,5 metros / 2 metros) dependiendo de la estancia y abarcando la mayor superficie posible.

Las dimensiones de las matrices generadas en el campus, aparte de no coincidir en el número de MAC´s registradas por cada aplicación, tampoco coincidían en el número de muestras. Esto fue debido a que 3 medidas procedentes de WIFI Analyzer fueron desestimadas por considerarse nulas.

Las nuevas medidas de las matrices fueron de 82 x 33 para Salamander y de 79 x 85 para el WIFI Analyzer.

De nuevo realizamos un entrenamiento de algoritmos con dichas matrices para observar como variaban los resultados respecto a la vivienda unifamiliar.

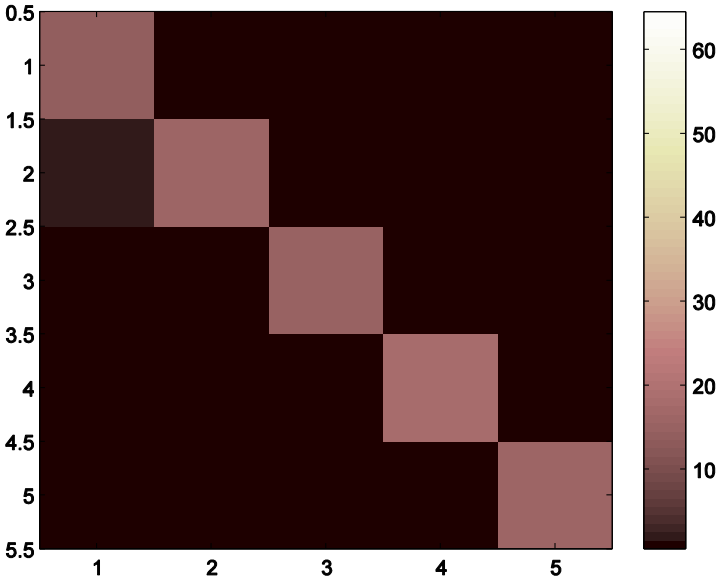
Algoritmos de clasificación	Datos de Salamander	Datos de WIFI Analyzer
KNN (K- Vecinos)	0,92	0,97
Naive Bayes	0,90	0,93
Multiclass SVM	0,95	0.96
Regresión multinomial	0,93	0,86

Figura 26. Tabla de acierto en clasificación de los algoritmos en el campus de la Universidad Carlos III. [8]

Podemos observar que ha mejorado de manera notable el porcentaje de acierto en la clasificación de todos los algoritmos respecto a las estancias de la vivienda unifamiliar.

Esto nos indica que nuestra herramienta de posicionamiento funciona mejor cuanto mayor es el número de puntos de acceso y más datos de potencia forman la matriz de entrenamiento de los algoritmos.

Vamos a observar de nuevo mediante una nueva matriz de confusión cuales son las muestras que generan más problemas a la hora de ser clasificadas. Como las tasas de acierto son muy aceptables para los 4 algoritmos, esta vez vamos a mostrar la matriz de confusión de MultiSVM para las muestras recogidas con aplicación de Salamander.



Salamander

Clase 1…….Hall Torres Quevedo
Clase 2…….Cafetería Puerta 1
Clase 3…….Cafetería Puerta 2
Clase 4…….Hall Agustín de Betancourt
Clase 5…….Reprografía

Para las estancias de la Uc3m no parece existir problemas de clasificación de las muestras. Las muestras pertenecientes a las estancias 1, 3, 4 y 5 han sido clasificadas al 100% correctamente. Solo ha existido algo de confusión con las muestras de de la estancia nº 2.

Figura 27. Mapa de confusión del algoritmo MulticlassSVM con datos recogidos mediante la aplicación Salamander. [8]

Esta nueva matriz de medidas de la universidad se adicionó a la matriz de datos de la vivienda unifamiliar, ya creada, para realizar un nuevo entrenamiento de los algoritmos de clasificación.

La dimensión de la nueva matriz se amplió hasta los 130 x 47 para Salamander y hasta los 127 x 99 para el WIFI Analyzer. De esta manera ampliamos las matrices en:

☑ Salamander → 5 estancias nuevas añadidas a las 4 anteriores, 82 muestras más añadidas a las 48 anteriores y 32 MAC´s nuevas añadidas a las 14 anteriores.

☑ WIFI Analyzer → 5 estancias nuevas añadidas a las 4 anteriores, 79 muestras más añadidas a las 48 anteriores y 84 MAC´s nuevas añadidas a las 14 anteriores.

Los resultados obtenidos para esta nueva matriz son:

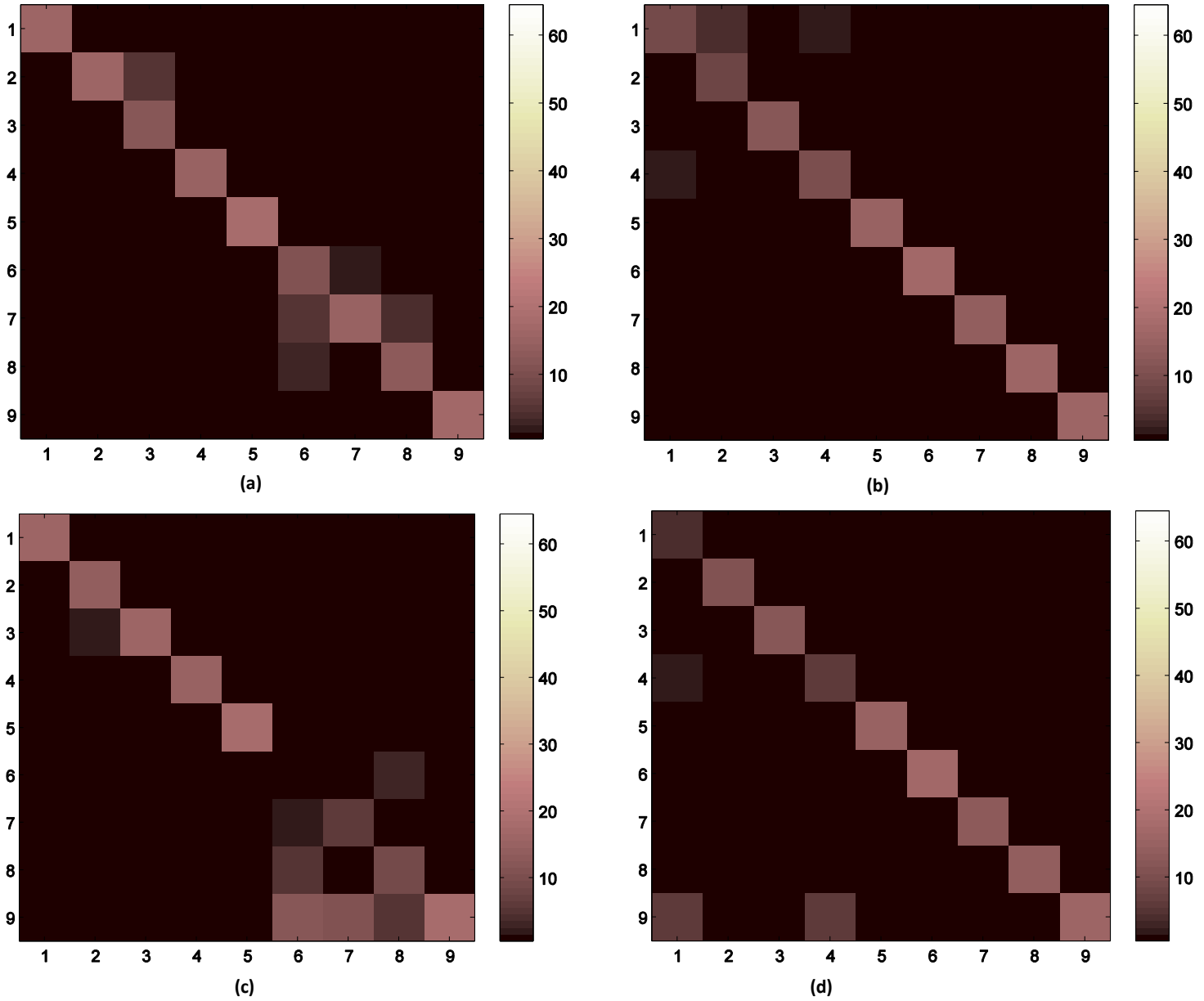
Algoritmos de clasificación	Datos de Salamander	Datos de WIFI Analyzer
KNN (K- Vecinos)	0,85	0,92
Naive Bayes	0,75	0,86
Multiclass SVM	0,71	0,85
Regresión Multinomial	0,79	0,55

Figura 28. Tabla de acierto en clasificación de los algoritmos en el caso mixto (Mezcla de datos). [8]

Estos resultados son los que ofrecen los algoritmos al ser entrenados con una matriz fruto de la fusión de las muestras de la vivienda unifamiliar y del campus de la universidad. Por separado hemos visto que respondían mejor en el campus de la universidad que en la vivienda unifamiliar. Pero nos interesaba saber cuál es el grado de repercusión que tendría en cada uno de ellos al tener que entrenarse y clasificar las muestras de esta nueva matriz.

La mayoría de ellos han mantenido una tasa de acierto bastante aceptable teniendo en cuenta que la tasa de acierto aleatoria es del 1/9.

Mediante unas nuevas matrices de confusión vamos a comprobar si los algoritmos han arrastrado problemas de clasificación de la vivienda unifamiliar y si han seguido manteniendo una tasa de acierto en clasificación alta para la Uc3m.



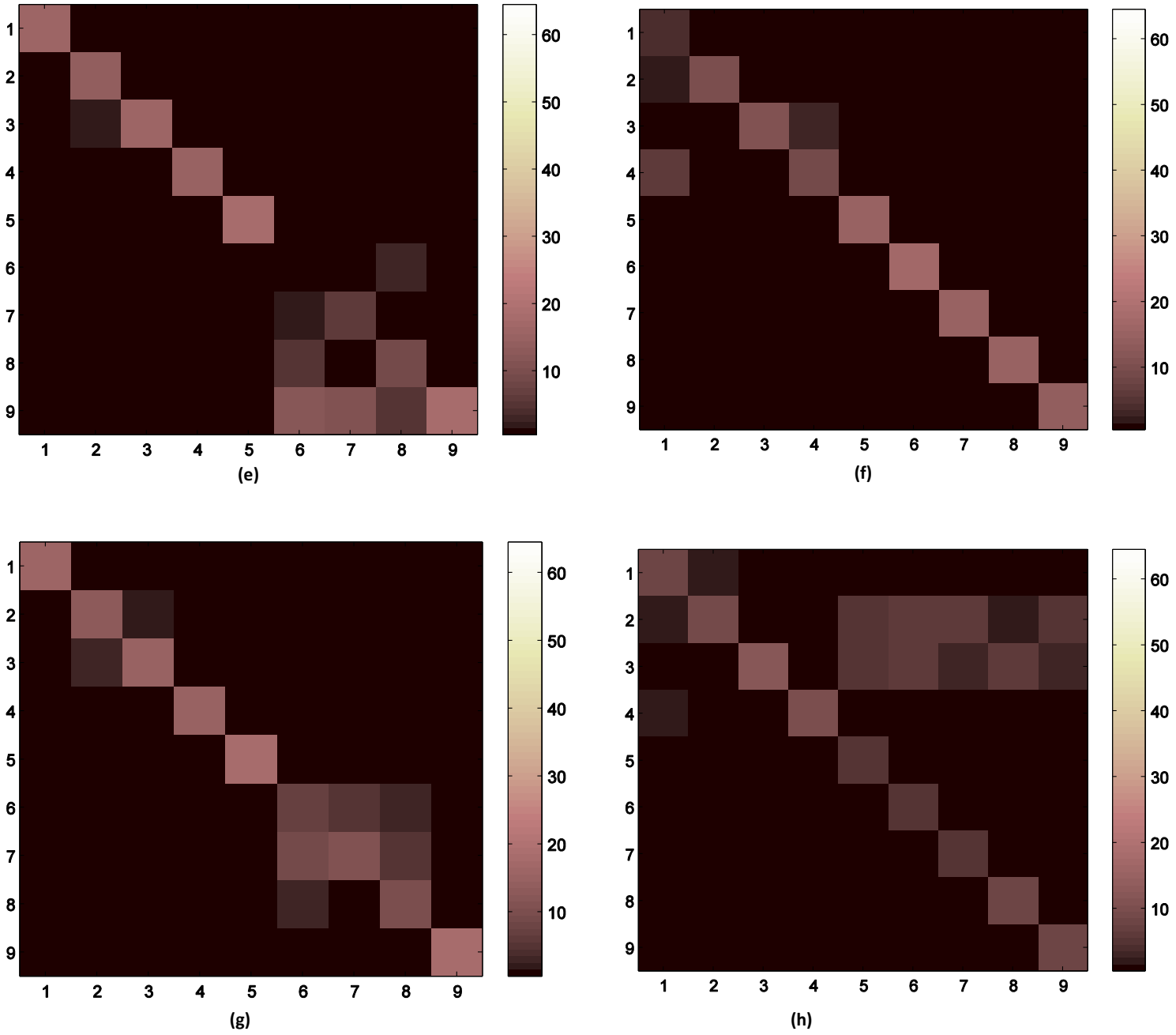


Figura 29. Conjunto de matrices de confusión de los 4 algoritmos para cada una de las aplicaciones usadas. [8]

Las matrices de confusión situadas a la izquierda (a, c, e, g) corresponden a la aplicación Salamander con los 4 algoritmos en el orden en que aparecen en la tabla. Las situadas a la derecha (b, d, f, h) corresponden a WIFI Analyzer con los mismos 4 algoritmos y en el mismo orden que aparecen en la tabla.

Antes de analizar las matrices debemos de tener en cuenta la clase que corresponde a cada estancia:

Salamander

Clase 1……..Hall Torres Quevedo
Clase 2……..Cafetería Puerta 1
Clase 3……..Cafetería Puerta 2
Clase 4……..Hall Agustín de Betancourt
Clase 5……..Reprografía
Clase 6……..Habitación 2
Clase 7……..Habitación 1
Clase 8……..Buhardilla
Clase 9……..Pasillo

WIFI Analyzer

Clase 1……..Habitación 1
Clase 2……..Buhardilla
Clase 3……..Pasillo
Clase 4……..Habitación 2
Clase 5……..Cafetería Puerta 1
Clase 6……..Hall Agustín de Betancourt
Clase 7……..Cafetería Puerta 2
Clase 8……..Hall Torres Quevedo
Clase 9……..Reprografía

Podemos observar que en la mayoría de las matrices de confusión todos los algoritmos arrastran los errores en clasificación, aunque de manera algo más suavizada, en muestras recopiladas en la vivienda individual y mantienen un porcentaje alto de clasificación para las muestras recopiladas en el campus de la universidad.

Todos los resultados que observamos en las tablas anteriormente propuestas has sido obtenidos mediante la configuración que los algoritmos traen por defecto. Pero ya mencionamos en la explicación de dichos algoritmos, estos nos ofrecían unas configuraciones distintas que también hemos decidido probar.

Mientras algunos algoritmos como el Naive Bayes apenas ofrecen una variación en el tipo de distribución (lineal o kernel), otros como el algoritmo KNN (K - Vecinos), tienen un amplio número de variaciones. Vamos, por lo tanto, a comprobar si cambiando los valores que el algoritmo presenta por defecto podemos obtener un mejor rendimiento.

Hemos hablado de la importancia de la elección del valor de K y de las distintas distancias que el algoritmo propone a la impuesta por defecto.

El valor elegido de K durante la obtención de resultados ha sido 5, pero debemos de comprobar si otros valores de K nos otorgarían mejores tasas de acierto.

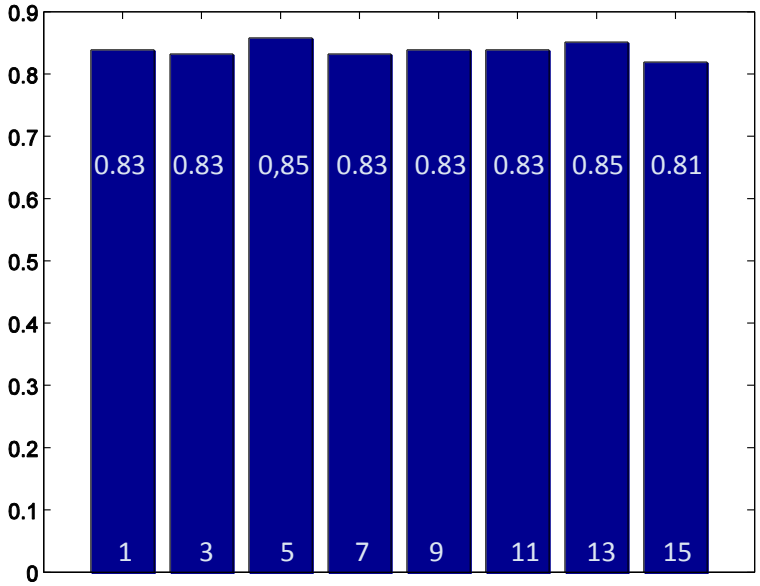


Figura 30.Valores de K para el algoritmo KNN. [8]

El valor que mejores resultados nos proporciona es el de K = 5, por lo que no es necesario volver a entrenar y realizar la clasificación de muestras con el algoritmo KNN con un valor de K distinto.

Otro factor que podemos modificar es la distancia utilizada por el algoritmo para medir la distancia entre los vecinos más próximos. Sabemos que por defecto usa la distancia euclídea, sin embargo probaremos con otros 3 tipos más que el algoritmo nos ofrece.

☑ Cityblock

☑ Cosine

☑ Correlation

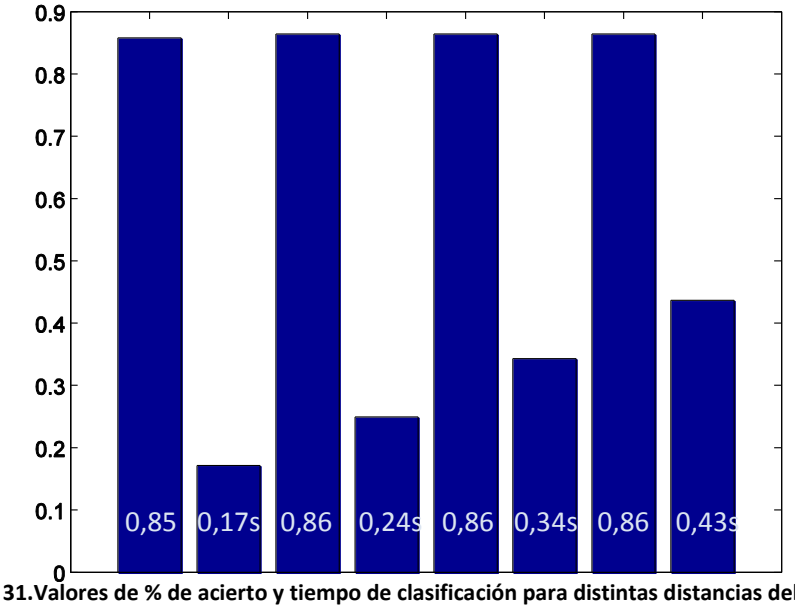


Figura 31.Valores de % de acierto y tiempo de clasificación para distintas distancias del algoritmo KNN. [8]

La primera de las columnas corresponde a la distancia por defecto y las otras 3 a las distancias cityblock, cosine y correlation respectivamente. Las tres distancias nuevas utilizadas aumentan la tasa de acierto del algoritmo, pasando del 85% al 86%, sin embargo aumenta progresivamente el tiempo de clasificación de las muestras (tiempo total), lo que descarta su elección. Posteriormente veremos un análisis completo de todos los tiempos de clasificación y entrenamiento de algoritmos.

Una vez hemos visto un caso en el que se refleja la importancia de un tiempo de clasificación bajo, vamos a mostrar dichos tiempos para todos los algoritmos empleados.

Para ello hemos calculado los tiempos de entrenamiento y clasificación de cada algoritmo. El tiempo de entrenamiento se refiere al tiempo que el algoritmo destina al aprendizaje de las muestras de la matriz. El tiempo de clasificación es el tiempo que el algoritmo emplea en analizar una muestra y decidir de qué estancia proviene.

Como los tiempos totales de entrenamiento y clasificación son distintos dependiendo del número de muestras, hemos realizado la media por muestra.

Tiempos(seg)	Entrenamiento	Clasificación	Entrenamiento	Clasificación
Algoritmos	Salamander	Salamander	WIFI Analyzer	WIFI Analyzer
KNN(KVecinos)	$3{,}78\text{x}10^{-3}$	$9{,}46\text{x}10^{-4}$	$2{,}65\text{x}10^{-3}$	$6{,}63\text{x}10^{-4}$
Naive Bayes	$3\text{x}10^{-2}$	$1{,}07\text{x}10^{-4}$	$16{,}6\text{x}10^{-3}$	$0{,}53\text{x}10^{-3}$
MulticlassSVM	9,43	$13{,}1\text{x}10^{-3}$	6,54	$5{,}65\text{x}10^{-3}$
R.Multinomial	22,13	$1{,}8\text{x}10^{-2}$	287,54	$1.29\text{x}10^{-4}$

Figura 32. Tabla de tiempos que emplean los algoritmos en el entrenamiento y clasificación de muestras. [8]

El tiempo de entrenamiento se corresponde a la totalidad de muestras ya que aquí no podemos hacer distinción entre una o varias muestras ya que si no entrenamos al algoritmo con todas las muestras de entrenamiento no podemos clasificar la muestra de test.

Observamos que la clasificación de las muestras es razonablemente rápida para todos los algoritmos. El problema está en el tiempo de entrenamiento, que hace que algunos algoritmos sean auto-descartados a pesar de ofrecer una buena tasa de acierto.

5. Evaluación de los resultados obtenidos.

Con todos los datos mostrados podemos realizar una evaluación de los resultados ofrecidos por los algoritmos para las dos aplicaciones empleadas en el diseño de la herramienta.

- **Salamander** → Los datos recopilados se pueden considerar más fiables al realizar 24 medidas por cada punto de muestreo en la estancia. De esta manera si existiesen picos de señal no reales, el error que estos introducen es menor. Por el contrario, esta aplicación tiene una sensibilidad menor que WIFI Analyzer detectando menor número de puntos de acceso en igualdad de condiciones. Esto es, situándonos en ambas recopilaciones de datos en idéntica posición.

  ☑ **KNN (K - Vecinos)** → Es el algoritmo que mejores resultados muestra en el caso de la vivienda unifamiliar y es el tercer algoritmo con mejores prestaciones en las estancias del campus de la Universidad. En el caso mixto destaca sobre el resto de los algoritmos situándose en primera posición.

  ☑ **Naive Bayes** → Está a la cola tanto en el caso de vivienda unifamiliar como en el campus de la Universidad, sin embargo, remonta una posición en el caso mixto. Sus prestaciones no son las más idóneas.

  ☑ **Multiclass SVM**→ Tiene un comportamiento desigual dependiendo de las estancias. En el caso de la vivienda unifamiliar es el segundo que mejores prestaciones da y en el campus de la Universidad es el mejor de ellos. Sin embargo, en el caso mixto, acusa mucho la mezcla de datos y es el último en la clasificación de porcentajes de acierto.

  ☑ **Regresión Multinomial** → Su comportamiento es muy homogéneo en los tres casos. Se coloca como segundo mejor clasificado, tercer mejor clasificado y repite la tercera posición en los casos de la vivienda unifamiliar, campus de la universidad y caso mixto respectivamente. Muestra ser un algoritmo que no acusa en exceso la mezcla de los datos de entrenamiento, pero sin alcanzar unas prestaciones idóneas en ninguno de los tres casos.

Ya conocemos los puestos que los algoritmos ocupan en las clasificaciones de eficiencia, clasificando una muestra y decidiendo la estancia de donde proviene. Ahora, debemos de hacer referencia al tiempo que emplea cada algoritmo en el entrenamiento y la clasificación de las muestras.

Si hablamos de tiempo de entrenamiento, el más rápido con mucha diferencia es el KNN y el que más rápido clasifica una muestra es el Naive Bayes seguido muy de cerca de nuevo por el KNN.

Por lo tanto observando los datos del comportamiento que tienen los algoritmos en cuanto a eficiencia y rapidez se refiere, cuando los datos para su clasificación y entrenamiento han sido recopilados con el Salamander, podemos decir que:

☑ **Vivienda unifamiliar** → El algoritmo que mejor se adapta a las condiciones que allí se dan es el KNN. Quizás si solo buscamos una rápida clasificación, y no nos importa el tiempo de entrenamiento, podemos decantarnos por el Naive Bayes, pero su pésima tasa de acierto no invita a su elección.

☑ **Campus de la Universidad**→ Aquí podríamos tener más dudas que en el primer caso. La mayoría de los algoritmos tienen una tasa de acierto bastante aceptable pero los tiempos de clasificación y entrenamiento varían demasiado. Quizás podríamos decantarnos por el Naive Bayes perdiendo algo de prestaciones en cuanto a la efectividad pero ganando en tiempo de clasificación y entrenamiento. Aquí dependería de lo que el usuario buscase.

☑ **Caso mixto** → Aquí de nuevo muestra su predominio el algoritmo KNN, con una diferencia de efectividad insalvable respecto al resto de sus competidores. Además, a su favor está el poco tiempo empleado en clasificación y entrenamiento.

- **WIFI Analyzer** → Los datos recopilados por esta aplicación pueden no ser en algunos casos del todo correctos. Esto es debido a que las capturas de datos son instantáneas y pueden colarse picos de señal no real en los mismos. Sin embargo, detecta un número mayor de puntos de acceso que Salamander lo que ayuda claramente a los algoritmos a aumentar su efectividad.

  ☑ **KNN (K - Vecinos)** → Es el algoritmo que mejores resultados muestra en los tres casos.

  ☑ **Naive Bayes** → Se comporta de manera muy homogénea en los tres casos. Es el tercero que mejores resultados presenta para la vivienda unifamiliar y repite puesto en el campus de la Universidad. Para el caso mixto asciende al segundo puesto.

☑ **Multiclass SVM** → Su progresión es inversa a la de Naive Bayes. Repite segundo puesto en los dos primeros casos, vivienda unifamiliar y campus de la Universidad, pero baja al tercer puesto en el caso mixto.

☑ **Regresión Multinomial** → Es el algoritmo que peor funciona para los tres casos repitiendo el último puesto en las 3 clasificaciones.

Ya conocemos los puestos que los algoritmos ocupan en las clasificaciones de eficiencia a la hora de clasificar una muestra y decidir la estancia de donde proviene. Ahora, debemos de hacer referencia al tiempo que emplea cada algoritmo en el entrenamiento y la clasificación de las muestras.

Si hablamos de tiempo de entrenamiento, de nuevo, el más rápido con mucha diferencia es el KNN y el que más rápido clasifica una muestra es el algoritmo de Regresión Multinomial seguido muy de cerca de nuevo por el KNN.

Por lo tanto observando los datos del comportamiento que tienen los algoritmos, cuando los datos para su clasificación y entrenamiento han sido recopilados con el WIFI Analyzer, podemos decir que:

☑ **Vivienda unifamiliar** → Se repite el mismo caso existente en la aplicación de Salamander. El algoritmo que mejor se adapta a las condiciones que allí se dan, es el KNN. De nuevo, la clasificación de muestras más rápida la realiza otro algoritmo, la Regresión Multinomial, pero su tasa de acierto y un excesivo tiempo de clasificación lo auto-descartan.

☑ **Campus de la Universidad**→ De nuevo el algoritmo KNN muestra su predominio ya que su efectividad es la mayor de todas y le acompaña el poco tiempo empleado en entrenamiento y clasificación. Elegir cualquier otro clasificador sería empeorar las condiciones.

☑ **Caso mixto** → Como todo parecía indicar, en el caso mixto también muestra su predominio el algoritmo KNN. El único apunte que podemos hacer es que el Naive Bayes sería una posible solución alternativa bastante fiable.

6. Desglose presupuestario.

En este apartado quedan reflejados los costes totales que ha tenido la elaboración de este Proyecto Fin de Grado. Dichos costes se resumen en la siguiente tabla.

Procesos del TFG	Número de horas destinadas.
Búsqueda de aplicaciones para recopilación de señal.	40 Horas
Recopilación de los datos de entrenamiento y test. Tratamiento de datos para trabajar con Matlab	25 Horas
Implementación del código Matlab para cada algoritmo empleado.	30 Horas
Entrenamiento de los algoritmos de clasificación y clasificación de las muestras de test.	40 Horas
Tutorías	15 Horas
Redacción del TFG	100 Horas
TOTAL	**250 Horas**

Figura 33. Tabla de tiempos empleados en la totalidad de los procesos del TFG. [8]

Hemos presentado ya las horas empleadas en el desarrollo del proyecto. Vamos a mostrar ahora el material que ha sido necesario emplear para realizar el proyecto y el coste que este tiene.

Material	Precio (Euros)
Ordenador portátil HP	426 Euros
Teléfono móvil HTC Desire.	300 Euros
Licencias de programas adquiridos.	200 Euros
Desplazamientos Vivienda – Universidad.	30 Euros
TOTAL	**956 Euros**

Figura 34.Tabla de costes registrados durante la realización del proyecto. [8]

Ahora debemos de poner un precio a la hora de trabajo para añadir a los costes materiales los costes personales.

Como aún no se posee el título ingeniero, vamos a tarificar nuestras horas como becario, a un precio estimado de 4,5 Euros/Hora.

Las horas dedicadas en las tutorías las tarificaremos aparte, y de manera doble. Serán horas tarificadas como becario, al estar yo presente, y también como horas empleadas por el director del TFG, mi tutor. Estas horas estarán tarificadas a 60 Euros/Hora.

Por lo tanto:

Horas empleadas como becario	240 Horas x 4,5 Euros/Hora = 1080 Euros
Horas empleadas por el director del TFG en tutorías.	15 Horas x 60 Euros/Hora = 900
Coste del material empleado	956 Euros
TOTAL	2936 Euros

Figura 35.Tabla de costes totales. [8]

7. Conclusiones.

Tras la realización del trabajo podemos extraer una serie de conclusiones del mismo:

☑ Hemos conseguido crear una herramienta de localización y guiado en interiores, mediante el uso de algoritmos de clasificación multiclase y ayudados por la herramienta matemática por excelencia, Matlab.

☑ Hemos realizado un estudio de varios algoritmos empleados para la clasificación multiclase de muestras de potencia de señal WIFI, exponiendo en cada caso cual es el más aconsejable.

☑ La herramienta creada carece de movilidad actualmente, pero está perfectamente diseñada para implementarla en un teléfono móvil. Sería la continuación a este trabajo.

☑ La consecución de las herramientas idóneas para la recopilación de muestras ha sido una de las partes más densa. Se realizó multitud de pruebas con varias aplicaciones distintas hasta conseguir la aplicación que cumplía nuestro objetivos.

☑ La implementación del código Matlab para observar el comportamiento de los algoritmos no ha sido sencilla. No eran problemas inabarcables, pero si multitud de detalles que ralentizaban la consecución de resultados.

☑ Una vez elegido el algoritmo clasificador a emplear, la herramienta se resume en dos pasos: entrenamiento y clasificación. Quizás el entrenamiento sea la parte más engorrosa y que más tiempo conlleva, pero la clasificación, dependiendo del algoritmo elegido, es prácticamente instantánea. Pero esto es lo que realmente buscamos, ya que el entrenamiento corre a cuenta de nosotros, y lo que queremos es que el proceso que concierne al usuario, la clasificación, sea lo más rápida posible.

Una de las alternativas que se proponen como continuación al trabajo es el traslado de esta herramienta a un terminal móvil, ya sea de la plataforma Android o IOs.

8. Referencias.

[1]. "Historia de las redes Inalámbricas", Universitat Politécnica de Valencia, Diciembre 2010.
Disponible en: http://histinf.blogs.upv.es/2010/12/02/historia-de-las-redes-inalambricas/

[2]. Tomás Simal, "Monográfico: redes WIFI" – Observatorio Tecnológico, Ministerio de Educación, Cultura y Deporte, Febrero 2011.
Disponible en: http://recursostic.educacion.es/observatorio/web/es/cajon-de-sastre/38-cajon-de-sastre/961-monografico-redes-wifi?showall=1

[3]. "Status of Project IEEE 802.11n"
Disponible en:
http://grouper.ieee.org/groups/802/11/Reports/tgn_update.htm

[4]. Jorge R.Rey, "El sistema de posicionamiento Global-GPS", University of Florida.
Disponible en: http://edis.ifas.ufl.edu/in657

[5]. Aroa Carcavilla Sanz, "Sistemas de posicionamiento basados en WIFI", Escola Politécnica Superior de Castelldefels, Febrero 2006.

[6]. "Ekahau Heat Mapper", Enero 2011.
Disponible en: http://www.webayunate.com/ekahau-heatmapper-visualiza-facilmente-los-lugares-con-mayor-cobertura-wi-fi/

[7]. "Qubulus, Indoor Positioning"
Disponible en: http://www.qubulus.com/

[8]. Ignacio Gallego Yuste. Elaboración propia.

[9]. "Wigle WIFI", Google Play.
Disponible en:
https://play.google.com/store/apps/details?id=net.wigle.wigleandroid&hl=es

[10]. Adolfo Chico Ciprián, "Diseño y desarrollo de un sistema de posicionamiento en Madrid, Diciembre 2009.

[11]. José Carlos López Díaz, "UN ALGORITMO GENÉTICO CON CODIFICACIÓN REAL PARA LA EVOLUCIÓN DE TRANSFORMACIONES LINEALES", Universidad Carlos III de Madrid, Julio 2010.

[12]. Constantino Malagón Luque, "Clasificadores Bayesianos, El algoritmo de Naive Bayes", Mayo 2003.

[13]. María Ángeles Dueñas Rodriguez, “Modelos de respuesta discreta en R y aplicación con datos reales.”, Universidad de Granada.

[14]. José Luis Alba Castro, “Curso de Doctorado: Decisión, estimación y clasificación", Universidad de Vigo.
Disponible en: http://www.gts.tsc.uvigo.es/~jalba/doctorado/SVM.pdf

[15]. http://www.uc3m.es/portal/page/portal/titulaciones_grado/primer_dia/mapa/mapa_lega

[16]. http://www.ekahau.com/products/heatmapper/overview.html

[17]. http://es.wikipedia.org/wiki/Direcci%C3%B3n_MAC

[18]. http://www.telepieza.com/wordpress/2008/05/08/redes-inalambricas-y-sus-conceptos-wifi-wireless-wlan-lan-wan-ssid-wep-wpa/

[19]. Harry Zhang, “The Optimality of Naive Bayes” University of New Brunswick, Canada. Disponible en:
http://www.cs.unb.ca/profs/hzhang/publications/FLAIRS04ZhangH.pdf